Genuine Sequential Estimation Procedures for Gamma Populations using Exact Evaluation Criteria

Except where reference is made to the work of others, the work described in this dissertation is my own or was done in collaboration with my advisory committee. This dissertation does not include proprietary or classified information.

Kevin Tolliver

Certificate of Approval:

Hyejin Shin
Assistant Professor
Mathematics and Statistics

Mark Carpenter, Chair
Associate Professor
Mathematics and Statistics

Peng Zeng
Assistant Professor
Mathematics and Statistics

George T. Flowers
Acting Dean
Graduate School

Genuine Sequential Estimation Procedures for Gamma Populations using Exact Evaluation Criteria

Kevin Tolliver

A Dissertation

Submitted to

the Graduate Faculty of

Auburn University

in Partial Fulfillment of the

Requirements for the

Degree of

Doctor of Philosophy

Auburn, Alabama
December 18, 2009

Genuine Sequential Estimation Procedures for Gamma Populations using Exact Evaluation Criteria

Kevin Tolliver

Signature of Author

Date of Graduation

Vita

Kevin Paul Tolliver, son of Kevin Leonadis Tolliver and Cheryl Gatewood Tolliver, was born in Washington, D.C., USA on April 17, 1985. He graduated from Edgewater High School in Orlando, FL in 2002. In 2005, he received his Bachelor of Science degree in Mathematics from Morehouse College in Atlanta, GA. He received his Doctor of Philosophy degree in Statistics from Auburn University in 2009.

Dissertation Abstract

Genuine Sequential Estimation Procedures for Gamma Populations using Exact Evaluation Criteria

Kevin Tolliver

Doctor of Philosophy, December 18, 2009
(B.S., Morehouse College, 2005)

80 Typed Pages

Directed by Mark Carpenter

In this dissertation, we develop genuine two-stage sequential procedures for bounded-risk and fixed-width confidence interval estimation for Gamma distributed populations, based on exact evaluation criteria. The term "genuine" refers to the fact that, in contrast to previous methods, the procedures proposed herein are based on the combined samples from both the first and second stages, rather than ignoring the data from the first-stage sample. Accordingly, the terminal sample size and the estimate are no longer independent, which complicates the theory development significantly. The term "exact" refers to the fact the procedures are not evaluated on asymptotic or large sample theory, as is common in the literature predating this dissertation, and the derivations are based only on the properties of the underlying distribution, i.e., Gamma. The practical application of each procedure was also considered and examples are given for both problems, i.e., bounded-risk and fixed-width.

Acknowledgments

I would like to thank my advisor, Dr. Carpenter, for his academic guidance. Under his direction I became more than a mathematician but a mathematical statistician. I am very appreciative of his encouragement and belief in my abilities when the material was not always clear. Without his support, this thesis would not have been completed. It has been a privilege working with him.

I also would like to extend my gratitude to the members of my committee: Dr. Peng Zeng and Dr. Hyejin Shin for their time and suggestions that led to me improving this work. Working with them in and out of the classroom has truly been a pleasure. In addition, I would like to thank Dr. Robert Norton for taking time out of his schedule to assist with the editing process.

I would also like to extend a special thanks to Kandace Noah for helping me understand how my research is applied in her field, and to Mary Mechler and Alicia Smith who gave important insight when writing my thesis.

I am extremely grateful to my family: my siblings, loving aunts, uncles, and grandmother. I give a special thats to my father, Kevin L. Tolliver, for being a positive role model in my life and my mother, Cheryl G. Tolliver, who always pushed me to do better myself.

Style manual or journal used Journal of Approximation Theory (together with the style known as "aums"). Bibliograpy follows van Leunen's *A Handbook for Scholars*.

Computer software used The document preparation package TEX (specifically LATEX) together with the departmental style-file `aums.sty`.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Statistical modeling is a technique used in many different scientific fields. By summarizing current results into one expression, statistical modeling aids researchers in explaining their current results. More importantly, observed outcomes could be utilized to make future predictions. In scientific experimentation there are many factors that can contribute to a certain outcome in a research experiment. A model simply refers to the outcome that is expressed as the mathematical function of these factors. In order to make these predictions, the data in these models are assumed to be random and have some underlying distribution where one or all parameters are unknown, and parameter estimation is used to fit these models.

The Gamma distribution is often assumed to be the underlying distribution to model right-skewed variables with positive support. Because of its flexibility this distribution has a wealth of applications and is often used to model random times-to-events. Two scientific fields of study where the Gamma distribution is most often used to model data are Survival and Reliability Analysis. In Survival Analysis, variables such as lifespans of organisms as well as time till a treatment takes effect can be modeled with the Gamma distribution. In Reliability Analysis Studies, lifespans of a system or systems components as well as chemical corrosion, e.g. can be modeled with the Gamma distribution. The information gained by statistical models in these two fields is used in developing life insurance plans, pertinent drug information, warranty information, quality control information, etc. A parameter often studied in these fields is Mean-Time-To-Failure (MTTF) that is very useful for systems used on a regular basis. A general queue also models times with a Gamma distribution. This is seen in various computer systems, call centers, and traffic flow management systems. Articles that use the Gamma distribution in modeling times relating to queues include: Choe and Shroff (1997), Amero and Bayarri (2007), Chu and Ke (1997), Clarke (1957). Though modeling times is the most frequent use, the Gamma distribution as a family of distributions can be assumed in any area where values have a positive support. For example, it is used frequently in climatology modeling both precipitation rates and precipitation intensity. This is seen in Maureil et. al (2007) and Gutowski et. al (2008). In addition, it is seen in censor imaging as shown in Chatelian (2007) and Chatelian (2008). These statistical models are reliant on the parameter estimation, therefore it is imperative for model fit that estimates under some criterion are accurate, low bias with low variation.

To have an accurate parameter estimate, Sequential Analysis is needed to determine how many observations are required An accuracy measure, such as standard error, is dependent upon the parameters of the unknown underlying distribution. In sequential analysis, all the observations are not sampled at once. In fact, having estimates with predetermined accuracy cannot be determined with a sample size known prior to sampling. It is well known that an unbiased estimator will become more accurate as the sample size increases. However, knowing the sample size needed to ensure the accuracy falls within the criterion is impossible to determine without any knowledge of the underlying distribution. Sequential problems such as assigned-accuracy problems deal more with the sample size than with the estimator itself. The final model estimates are dependent upon information gained in prior sampling.

Historically researchers have calculated measures of accuracy for sample design based on incorrect assumptions about the underlying distribution. For many years, the underlying distribution for the data is assumed to be Normal, even for time estimates where there is a positive support and the data is right skewed. Sampling that assumes that the data is Normal when it is not introduces the risk of not actually meeting the criteria. It could also lead to sampling more observations than is needed to meet the specified criterion. This is a very prevalent problem in statistics since many experiments and surveys are restricted by budgetary restraints.

1.2 Research Question

Our focus is on developing a sampling procedure that will ensure an accurate estimator, which means that the estimator will have a low bias and low variation. The model assumes that the estimator is unbiased so the concern is restricting the variation. This dissertation looks at two different problems involving predetermined accuracy. The first problem is to ensure that the risk falls within a bound and the second problem is to ensure the width of the interval estimator is within a bound. In doing so, we will completely avoid using large sampling theory. There will be no asymptotic approximation of the underlying distribution and because this is done our procedure will hold for any number of initial observations.

We develop the mathematical theory that ensures that the risk is within a pre-specified bound under a genuine two-stage sampling scheme that assumes that the data comes from a Gamma population. The term “genuine” refers to the fact that the sequential procedure is based on the combined sample from both stages. It may seem fairly obvious that a genuine two-stage estimation procedure will yield better results than one that disregards one of the two samples. This has not always been implemented. Arriving at a two-stage procedure that ensures risk is within a bound may come with a cost and could result in sampling more observations than previous sampling procedures. A more practical problem is considered where the goal is to sample the fewest number of observations that achieve this goal to avoid oversampling as described by Wald (1947). Using a relationship between risk and interval estimation, a genuine two-stage fixed-width interval estimator sampling scheme is produced that is unlike anything that has ever been done in this field before.

1.3 Sequential Analysis and Multistage Designs

Sequential analysis is a statistical theory of data where the final sample size is not known prior to the study. Sampling procedures where the final sample size is known prior to sampling is known as a fixed-sample size procedure. Sequential sampling procedures are used over fixed-sample size procedures for (1) ethical reasons, (2) conceivability reasons, and (3) economical reasons. For example, in a drug trial for reducing hypertension, if there are m initial observations where some of them develop side effects or there is significant evidence that the true mean is low, then medical ethics forbid further sampling. With other instances, arriving at an alternative solution is inconceivable. An example of conceivability reasons considers an industrial process. There is no known way of determining when a process will become out of control with a fixed-sample size. There are occasions when sequential analysis is economical. An example of this is any sampling where there is an attached cost to each observation. Sequential analysis can reduce the number of observations, which will consequently reduce the cost of the experiment. Finding assigned accuracy estimators for parameters of a Gamma process or population can be all three. As noted before, the Gamma is often assumed in modeling times in clinical trials. There is no conceivable solution for determining the final sample size needed to achieve predetermined accuracy with a fixed sample size. Since one objective is to sample the fewest number of observations that achieve a certain goal, it has economical applications. For all of those reasons, a sequential design needs to be implemented to achieve pre-assigned accuracy.

Sequential analysis consists of two components: (1) the *stopping rule* and the *decision rule*. The stopping rule indicates whether or not sampling should be stopped after m observations or whether additional observations should be sampled. A stopping rule is characterized as a mechanism for deciding whether to continue or stop a process on the basis of the present position and past events, which will almost always lead to a decision to stop at some time. The final resultant sample size N is called the terminal sample size. The decision rule tells what actions need to be taken after sampling has been stopped.

Definition 1.1 *If m is a known predetermined sample size, a sample is said to be sequential if the terminal sample size N is not fixed, i.e.*

$$P(N = m) < 1 \quad for \;\; m, N \in \mathbf{N}.$$

The emphasis in this context is on having an estimator that will fall below some predetermined accuracy, the terminal sample size N will be the final sample size that ensures this is the case. The terminal sample size depends on earlier observed information, $X_1, X_2, ..., X_m$ making it a random variable.

The optimal sample size (n^*) is the number of observations that best achieves a researchers goal. This can mean a number of things for different problems. In our context, the optimal sample size is the fewest number of observations that ensures that our estimator is accurate under some predetermined criteria. The optimal sample size is fixed and is dependent upon the unknown parameters of the underlying distribution. In an ideal situation the terminal sample size will equal that of the optimal sample size. The terminal sample size is assessed by looking at the ratio of the expectations,

$$E[N/n^*]. \tag{1.3.1}$$

The performance of the terminal sample size can be evaluated asymptotically

$$\lim_{m \to \infty} E[N/n^*] = 1. \tag{1.3.2}$$

and

$$\lim_{m \to \infty} Var[N/n^*] = 0. \tag{1.3.3}$$

In sequential analysis, there are two subfields: (1) purely sequential designs and (2) multistage designs. Earlier it was mentioned that the final sample size is not known prior to the start of sampling. However, this does not mean that each observation is observed one at a time. With purely sequential designs, each observation is observed one at a time and an analysis is performed after each observation is drawn. Whereas in multistage designs, multiple observations are drawn at a time,called a stage, and there is a cap on the total number of stages. The terminal sample size does not necessarily consist of the prior m observations. In some instances the m observations are used to determine the terminal sample size and then disregarded for the analysis.

Definition 1.2 *Let* $\mathbf{X}_1$=$\{X_1, ..., X_{m_1}\}$ *be an initial sample. For a decision rule* δ *subsequent samples are* $\mathbf{X}_i$=$\{X_{m_{i-1}+1}, ..., X_{m_i}\}$ *or* $\mathbf{X}_i = \emptyset$ *for* $1 \leq i \leq k$. *The sample* $\mathbf{X} = \cup_{i=1}^{k}\mathbf{X}_i$. *is a genuine k-stage sample.*

The advantages of purely sequential problems are that they yield better statistical results and the procedure will have a reduced chance of over sampling as described by Wald. However depending on the design, multistage sampling can be more cost efficient and more manageable.

For example, consider the problem of determining when an industrial process becomes out of control; data is read after each observation. In such cases, it will be practical to use a purely sequential design. As the data is read, the process can immediately determine when it has become out of control and there is no need to continue sampling. However in a clinical trial, it is not practical to treat one subject at a time. A multistage design is needed.

Multistage problems are currently used in a wide range of areas. Some multistage sampling schemes use a set of observations from the population as their initial sample. The subsequent samples consist of analyzing a subset of that initial set. This is seen in the U.S. Census' Current Population Survey multistage method, given in Moore, McCabe, and Craig (2007). This is also seen in crop management, Finney (1984), as well as multistage cluster analysis Phillipi(2005). In the context of modeling times, multistage designs are often used in adaptive designs. In a broad overview of adaptive designs, several examples were given where statistical procedures were modified during the conduct of clinical trials. It is not only efficient to identify clinical benefits of the test treatment under investigation, but also to increase the probability of success of clinical development, Chow and Chang (2008). Most adaptive designs in clinical trials can be referred to as adaptive randomization and group sequential designs. With the flexibility for stopping a trial early due to safety, futility and/or efficacy and sample size re-estimation at interim for achieving the desired statistical power. In an article on unified theory of two-stage adaptive designs the mathematical theory is proposed to adaptations in literature. To summarize, the adaptations alter the sampling distribution, which means the assumed results may not be true, Lui, Proschan, and Pledger (2002). For example, for two-stage adaptive tests in particular, changes in the sampling distribution can occur. Only recently it has been thought of to alter the p-value. In their article, they arrive at a number of useful theories on two-stage adaptive designs.

Using a large number of stages makes organization more complicated and both administrative expenses and interest charges on the large investment increase. This is also the case with a number of other sequential sampling procedures. Because of this, in literature there are many two and three-stage designs; Mukhopadhyay and Pepe (2006), Mukhopadhyay and Zacks (2006), Yao and Venkatraman (1998), Satagopan et al. (2002), Whittemore (1997), Jinn et al. (1987), Lorden (1983), Mukhopadyay (1995), etc., and not as many four, five, and six-stage designs.

Squared error loss is a measure of an estimate's distance from its true parameter.

Definition 1.3 *If $A > 0$ is constant specified by the experimenter that penalizes deviations more or less as need be, squared error loss of an estimator with n observations is the squared distance between a parameter θ and its estimator $\hat{\theta}$:*

$$L_n(\theta, \hat{\theta}) = A(\theta - \hat{\theta})^2.$$

In practice, this measure is assessed by its expected value, called *risk*. The risk gives an indication of the reliability of an estimate. High risk indicates the estimator is unreliable while a low risk indicates the estimator is reliable. Increasing the sample size is an action taken to lower risk. One method of accuracy measure in sequential analysis is the bounded risk estimator.

Definition 1.4 *For a predetermined risk bound w, a bounded risk estimator with the terminal sample size N number of observations is the expectation of the squared error loss*

$$R_N(\theta, \hat{\theta}) = E(L_N) \leq w.$$

It is well documented that the risk will be a multiple of the variance plus the bias of the estimator squared. Consider the Normal distribution with known variance; in this instance, bounding the risk is an easy calculation. However, if the mean and variance are unknown, then this problem cannot be solved with a fixed sample size. With fitting statistical models with parameter estimation, the goal is to have accurate mean parameter estimates.

Another accuracy measure in sequential analysis is the fixed-width interval estimator.

Definition 1.5 *For a predetermined width d, a $1 - a$ fixed-width interval estimator of a real-valued parameter θ is any pair of functions $L(X)$ and $U(X)$, with $L(X) < U(X)$ for with the inference $L(X) < \theta < U(X)$ is made. We say C_X is the interval $[L(X), U(X)]$, The width of the confidence interval is simply $U(X) - L(X) \leq d$, and $P(\theta \in C_X) \geq 1 - a$.*

Fixed-width confidence intervals are a large part of sequential estimation. It is known that interval estimation is more informative than point estimation due to the $P(\hat{\theta} = \theta) = 0$ for any continuous distribution. Interval estimators consist of width of the interval and the coverage probability. There are merits to both. A small coverage probability implies that the researcher has a larger chance of making an error, whereas a large width is uninformative.

1.4 The Gamma Distribution

The model assumption for the proposed sampling scheme is that the underlying distribution is Gamma. The focus is on estimating the mean parameter of the Gamma distribution. The Gamma distribution is a flexible right-skewed distribution that has a variety of applications. It is often used in modeling times-to-event that is seen in biological science, engineering, ecological, and probability fields. The density is

$$f(x) = \frac{1}{\Gamma(\alpha)\lambda^{\alpha}} x^{\alpha-1} e^{-x/\lambda}, \ \ for \ \ x > 0, \tag{1.4.1}$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ and $\alpha, \lambda > 0$, with $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ and $\alpha\Gamma(\alpha) = \Gamma(\alpha+1)$. Note, α and λ are referred to as the shape and scale parameters, respectively. A property with the Gamma density is that it is closed under scalar product. That is if $X \sim Gamma(\alpha, \lambda)$, then

$$Y = cX \sim Gamma(\alpha, c\lambda). \tag{1.4.2}$$

The sum of k Gamma random variables with shape α and scale λ is

$$\sum_{i=1}^{k} X_i \sim Gamma(k\alpha, \lambda). \tag{1.4.3}$$

The moment generating function of this distribution is

$$M_X(t) = \left(\frac{1}{1-\lambda t}\right)^{\alpha}, t < 1/\lambda. \tag{1.4.4}$$

Which makes the mean and variance

$$EX = \alpha\lambda \ \ \ and \ \ \ Var(X) = \alpha\lambda^2. \tag{1.4.5}$$

1.4.1 Special Cases of Gamma Distribution

Some special cases of the Gamma distribution will be noted as they are referenced throughout this dissertation. Suppose X is distributed with Gamma with shape α and scale λ. If the shape parameter α is one, then X is exponentially distributed with scale λ. It is well documented that adding k exponentially distributed variables will yield a Gamma distribution with shape k and scale λ, as noted in equation (1.4.3). If the shape parameter is an integer then the variable is Erlang with shape α and scale λ. If the scale is two, then X becomes a Chi-Square distribution with parameter 2α. It follows by (1.4.2)

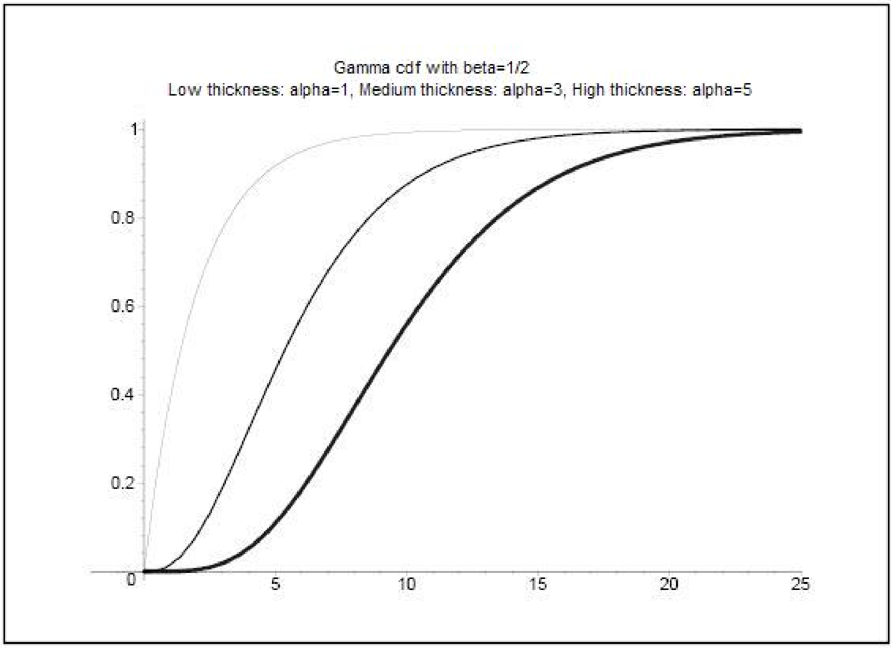


Figure 1.1: Gamma CDF

that $2X/\lambda \sim Chi - Square(2\alpha)$. Finally if there are two independent Gamma distributed variables, $X \sim Gamma(\alpha, \lambda)$ and $Y \sim Gamma(\beta, \lambda)$, then $\frac{X}{X+Y} \sim Beta(\alpha, \beta)$.

1.4.2 Estimation

As previously stated, the Gamma distribution is widely used in engineering, probability, ecological and biological science fields. The problem of finding reliable estimators for the mean dates back to the early 1950s. There are a number of different methods that can be used in finding estimators for this distribution: method of moments, maximum likelihood, and least squares. In particular, for this dissertation, the maximum likelihood method is used to obtain the estimator for λ and the method of moments estimator is used to obtain the estimator for α

The maximum likelihood method of estimation is the most popular technique for deriving estimators. This technique has many ideal properties including the fact that it yields the best unbiased estimators. Using (1.4.1), the likelihood function for n identically and

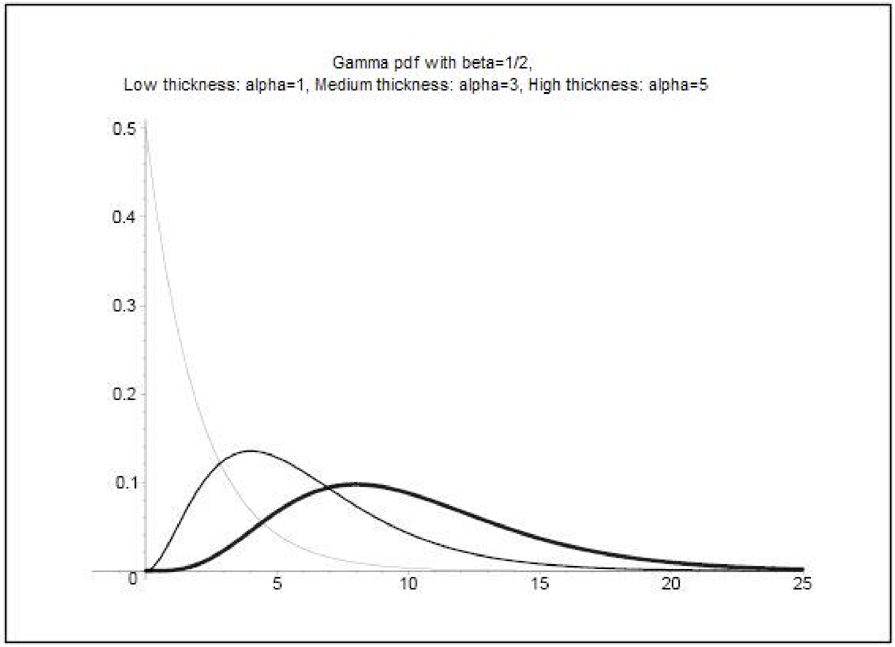


Figure 1.2: Gamma PDF

independently distributed variables. The likelihood function becomes:

$$L(\alpha, \lambda) = \frac{1}{[\lambda^{\alpha}\Gamma(\alpha)]^n} \prod_{i=1}^{n} x_i^{\alpha-1} e^{-1/\lambda \sum_{i=1}^{n} x_i} \quad x_i > 0,\ for\ i = 1, ..., n$$

To ease computation, the natural logarithm of the likelihood is taken. This can be done because the natural log function is a monotone function, so the likelihood will maintain its optimum values. If shape is known, the maximum likelihood estimator for λ can be easily obtained and is shown to be

$$\hat{\lambda} = \bar{X}/\alpha. \tag{1.4.5}$$

If the shape is unknown, no close form maximum likelihood estimator or numerical solution needs to be given to arrive at its maximum likelihood approximation. This is the reason the maximum likelihood estimation approach is not used for the shape parameter.

The method of moments estimator is another common method for estimating a parameter. It works by setting the k^{th} moment to the sum of x_i to the k^{th} power,

$$\hat{E}(X^k) = \frac{1}{n}\sum_{i=1}^{n} X_i^k \quad i = 1, ..., n \quad k \in \mathbb{N}.$$

Shape parameter (α), which is widely considered a nuisance parameter. Because of the unattainability of a best unbiased estimator, method of moments estimator is used,

$$\hat{\alpha} = \bar{X}^2/S^2. \tag{1.4.6}$$

This is a slightly biased estimate that is asymptotically consistent.

1.5 Literature Review

Early elements of sequential analysis appear in the 17th and 18th century, when mathematicians Huyghens, Bernoulli, Montmort, DeMoiver, LaGrange and LaPlace worked on the Gamblers Ruin problem, (Ghosh and Sen 1991). This famous probability problem tries to determine at what point gamblers will completely deplete their funds. Dodge and Romig in 1929 were the earliest to apply what is now known as sequential analysis to a statistical problem. They developed a double sampling test, where two samples were taken and the proportion of defective units was observed. Shewart in 1931 developed theory on what instance does an industrial process become out of control. Wald in 1947 produced a well known book on sequential analysis that sparked interest from several authors world wide.

1.5.1 Bounded Risk Estimation

A common sequential problem dealing with preassigned accuracy is bounded risk estimation. For populations with known variance, there is a fixed-sample size solution; no sequential methods need to be implemented. The problem arises when nothing is known about the population. Stein (1945) proposed a two-stage bounded risk estimation procedure for Normal populations. This procedure incorporated using the standard deviation from the initial sample to yield the proper terminal sample size. Modeling times with a Normal underlying distribution will not yield ideal results because times are often skewed to the right. A better distribution to assume when modeling times is the Exponential distribution.

Birnbaun and Healy (1960) developed a two-stage bounded risk estimation procedure for Exponential processes. This sampling scheme assumed that the underlying distribution was Exponential and found ways to bound the scale parameter using Chi-Square transformations. Their result can be summarized as follows: if there are $m \geq 3$ initial observations and

$$B_{BH} = \frac{Am^2}{(m-1)(m-2)}$$

then

$$N_{BH} = \lceil \frac{B_{BH}\bar{X}_m^2}{w} \rceil \tag{1.5.1}$$

is the sample size required so that the risk of the estimator is within the bound w. However, this procedure is not a genuine two-stage sampling procedure. The solution is only based on the second sample. The initial sample is used to determine the sample size required in achieving bounded risk and then it is not included in the final estimate. This is done because taking observations from two different samples alters the sampling distribution. It is true that bounding the risk cannot be done with a fixed sample size. However, disregarding readily available information is wasteful. This concept was improved by adding observations from the initial sample to the second sample, thus making the procedure a genuine two-stage sampling procedure. Although, this proved to be an asymptotically great bounded risk-estimator (Kubokawa 1989), no actual proof was provided for this result and it is uncertain if it holds for any number initial observations (> 3).

Various works in sequential estimation of scale parameter of the Exponential distribution is done by Mukhopadhyay (1995), (2006), (2006a), (2006b), (2007), etc. Mukhopadhyay and Pepe (2006) record an exact genuine two-stage sampling procedure. Their result which holds for any initial sample size greater than three, can be summarized as follows: if there are $m \geq 3$ initial observations and

$$B_{MP} = \frac{2Am(m+1)}{(m-1)(m-2)}$$

then

$$N_{MP} = \lceil \frac{B_{BH}\bar{X}_m^2}{w} \rceil \tag{1.5.2}$$

is the sample size required so that the risk of the estimator is within the bound w. The consequence of this is the expected value of the terminal sample size is more than twice that of the initial sample size; meaning on average the researcher will sample more than twice the observations needed. This is referred to as a penalty for exact bounded risk estimation. Exploring the distribution of this terminal sample size, a reduction of this terminal sample size could be found making this exact procedure more practical, Zacks and Mukhopadhyay (2006).

1.5.2 Fixed Width Interval Estimation

The next sequential problem is restricting the interval estimators width. Interval estimation is one of the fundamental aspects of statistics. Presenting interval estimators are often preferred over measures of variation, such as risk. This is probably due to the fact that confidence intervals can yield better interpretations. This is particularly important when the estimate does not have a Normal sampling distribution, Ramsey and Shafer (2002). Both measures give flexibility to the estimator, but interval estimators give results in terms of what is probable, i.e. the probability that the true parameter lies within a $1-a$ confidence region is $1-a$.

Fixed-width confidence consists of relatively high probabilities and relatively narrow interval widths. Traditionally, they are of the form

$$C_X = \{\theta | \bar{X}_N - d < \theta < \bar{X}_N + d\}, \tag{1.5.3}$$

with the terminal sample size N. There is no fixed sample size solution to this when the variance of a distribution is unknown. The width of the interval estimator is dependent upon the variance of the distribution. Stein (1949) solves the Normal fixed width problem by proposing a two-stage procedure to bound the confidence interval for mean μ when variance σ^2 is not known. The terminal sample size of this procedure is given below,

$$N = \max\{m, \lceil \frac{b^2_{m-1,1-a/2} S^2_m}{d^2} \rceil\}, \tag{1.5.4}$$

where $b_{m-1,1-a/2}$ is the $1-a/2$ point of a t with $m-1$ degrees of freedom. This uses the fact that $b_{m-1,1-a/2}$ will be larger than $z_{1-a/2}$. This procedure was shown to be asymptotically inconsistent. As the initial sample size gets large, the ratio between widths of this procedure's sample size and the optimal sample size will be $(b_{m-1,1-a/2}/z_{1-a/2})$. Ghurye (1958) proposes a two-stage fixed-width confidence interval for a location parameter of a general density, $f(x)$, along the lines of Stein. This is not used for mean. Chow and Robbins (1965) record a purely sequential interval estimator for the mean of a general density $f(x)$. This sequential result uses an initial sample of m observations, then chooses the first n for which the following is achieved size is

$$N = \min\{n \geq m | n \geq d^{-2} z^2_{a/2}(S^2_n + n^{-1})\}, \tag{1.5.5}$$

However, their procedure uses asymptotic theory. This is not a practical approach for model estimates because it observes observations one at a time and we are avoiding Normal approximation. A general method for determining fixed width confidence intervals is given by Khan (1969). This method like the previous methods use Normal theory; in it he discusses almost sure convergence, asymptotic consistency, and asymptotic efficiency. Research is continuing to be developed in this area. For example, Mukhopadhyay, Silva, and Waikar (2006) develop a two-stage sampling procedure to which they compare Steins fixed width interval approach (1949) and Chapmans fixed width interval approach (1950).

As noted before when estimating mean time, it is better to model with the Exponential distribution. Govindarajulu (1995) developed a sequential estimator for the mean of an Exponentially distributed population. This result is more applicable to modeling times, it is summarized that as follows: if

$$z_n = z[1 + n^{-1}(1 + z^2)/4 + o(n^{-1})]$$

$$N_G = \min\{n \geq m | n \geq z_n^2 \bar{X}_n^2 / d^2\} \tag{1.5.6}$$

will bound the risk. This procedure again uses Normal approximation.

However because of the shape of this distribution, no research has been found on restricting the width of the interval without using asymptotic approximation.

1.5.3 Modeling Times with Gamma

It should be noted that both of the prior subsections ended with sequential research in statistical modeling for Exponential populations. This is because there is not much research in this area for Gamma populations. However, in the same instances where the Exponential model can be used, so can the Gamma. Specifically statistically modeling random times, such as mean time-to-failures, are assumed to be Exponential. Where the longer one survives the smaller the probability is for continual survival. This is not always the case. For example, it is charted for life expectancy of an infant that there are many casualties during the first few months of birth. So for a short period of time, the life expectancy increases the longer the infant lives. It will be more appropriate to model infant life expectancy with a Gamma or Weibull distribution. Both of these distributions are more general forms of the Exponential distribution. There are several works that discuss modeling MTTF as a

Gamma distributed variable: Coit and Jin (1999), Shapiro and Wardrop (1978), Barber and Jennison (2002), etc. These articles provide examples of when the Gamma should be used over the Exponential distribution. For example, when modeling failure times with a known number of failures and missing values are present. The time between one failure and the last record is Gamma with known shape. This happens often when data is recorded periodically and not after each failure, Coit and Jin (1999).

1.6 Dissertation Layout

In the second chapter, two-stage bounded risk estimators are developed. The performances of these sample sizes are evaluated through simulation. Use of numerical methods is implemented to reduce the value for the sample size, making the estimator more asymptotically consistent. In the third chapter, a fixed-width interval estimator is created, and another example is given to illustrate how it works and how it relates to queueing theory. The final chapter summarizes the results of this dissertation and discusses future research problems in this area.

Chapter 2
Bounded Risk Estimation

The goal for the bounded risk problem is to sample the fewest number of observations so that the risk is just within a predetermined bound. Birnbaum-Healy (1960) developed a two-stage sampling procedure for the Exponential distribution; however their method does not use the information obtained from their initial sample in their final estimate. Mukhopadhyay and Pepe (2006) develop a two-stage sampling procedure for the Exponential distribution that combines the initial sample with the second sample to derive the final estimate. In this chapter, we generalize Mukhopadhyay and Pepe's result to the Gamma distributed populations. It should follow that when the shape is equal to one, our results will be exactly that of Mukhopadhyay and Pepe.

Additionally, we introduce some notations and basic concepts of decision theory as it applies to the Gamma distribution. First, the risk bounds are found when only the shape parameter is known. Secondly, risk bounds are found when both parameters are unknown. We also evaluate the performance of our bounds theoretically and through simulations and make possible improvements.

2.1 Shape Known and Scale Unknown

There are a number of reasons the shape known case is studied: (1) There are particular instances where the shape parameter is either known or can be assumed as known, (2) studying the alpha known case allows us to see how robust the Exponential assumption is, (3) there are times when the shape parameter is not known but there is a mathematical theory that allows us to assume the shape parameter is known, and (4) it lays the ground work for when shape parameter is unknown.

MTTF is often estimated as an Exponential random variable. This is merely a special case of the Gamma distribution when the shape is known and equal to one. Mukhopadhyay discusses this in a number of articles (1995), (2006), (2006a), (2006b), (2007), etc. However in many cases, MTTF is modeled with a Gamma distribution when the shape is known and not equal to one. Dopke (1992) and Coit and Jin (1999) discuss estimation of the MTTF as a Gamma random variable. The example given by Coit and Jin is when the time between each failure is not recorded. If there are k failures in a span t, then the MTTF is Gamma distributed with known shape k and unknown scale. They elaborate on why each failure time is not always recorded saying, "this is understandable because the elapsed time meter records time for the entire assembled item and not the individual components." This is similar to the idea of sum of Poisson process random variables. Other examples when the shape is known occur with modeling times and Normal distribution; there are modeling times when the shape parameter is assumed known just as there are instances in the Normal distribution when variance is assumed known. This can happen for a number of reasons; either there is so much historical evidence that the shape is consistent, there exists some mathematical theory for the shapes value, or the actual shape is of little concern as long as it is within reason. For example, in Maurellies (1999) precipitation models, he discusses the actual unimportance of knowing the exact shape. They state that since the data is right skewed, it is important to model the data with a low shape value. In this dissertation, they simply model precipitation intensity with $\alpha = 2$. In each of these examples it is important to have reliable estimates.

This is not the only reason for exploring the shape known, scale unknown case. Studying the shape known case also gives an idea of how robust the assumption of an Exponential distribution is. Mukhopadhyay and Pepe's (2006) result is only for the Exponential distribution. If there is some uncertainty that the shape is one, then there is no validity to their procedure. These forementioned reasons provide justification for studying the shape known case.

Our goal is to develop a reliable estimation sampling scheme for when only the scale is unknown. If the shape α is known and it is only desired to estimate the scale λ, our goal is to find the fewest number of observations n that will make its associated *risk* function less than or equal to a predetermined risk $w > 0$. Recall the mean of a Gamma distribution with parameters α and λ is $\alpha\lambda$. Hence the risk is $R_n(\alpha\lambda, \hat{\alpha\lambda}) \leq w$. If the shape is known the problem reduces down to:

$$R_n = \frac{A\lambda^2}{\alpha n} \leq w. \tag{2.1.2}$$

This implies, $n \geq \frac{A\lambda^2}{\alpha w}$. Let the optimal sample size:

$$n^* = \lceil \frac{A\lambda^2}{\alpha w} \rceil. \tag{2.1.3}$$

This guarantees an integer value, which will ensure that the risk is within our bound. Sampling more observations than n^* is considered oversampling, sampling fewer observations than n^* will yield a high risk and thus an unreliable estimate.

Notice n^* is dependent on the unknown parameter λ, so a sequential sampling procedure must be implemented to ensure that knowledge can be gained on this parameter. A pilot sample of m observations $X_1, ..., X_m$ i.i.d variables will be taken following a Gamma distribution (α, λ), with $m\alpha > 3$. From this sample the maximum likelihood estimator of λ can be found using the maximum likelihood estimator $\overline{X}_m/\alpha$, see (1.4.5) to see how this was derived. That estimate is used to determine the terminal sample size N. This quantity will guarantee that we do not exceed the necessary number of observations for the statistical procedure by too much, as it might be costly or impractical, yet not fall short of an appropriate sample size either. After observing the first m observations, our first stage, a decision is needed to determine if the procedure can continue with the m observations, or if more need to be added, our second stage. Yielding our two-stage procedure.

Theorem 2.1 *If $X_1, ..., X_m$ i.i.d. Gamma (α, λ) initial observations are drawn, $(m\alpha \geq 3)$ and B is chosen to be*

$$B = A\alpha^2 \left[\frac{m}{\alpha} - \frac{2m^2\Gamma(m\alpha - 1)}{\Gamma(m\alpha)} + \frac{(m^3\alpha + m^2)[\Gamma(m\alpha - 2)]}{\Gamma(m\alpha)} \right]. \tag{2.1.5}$$

and the terminal sample size is chosen to be

$$N = max \left(m, \lceil \frac{B\overline{X}_m^2}{\alpha^3 w} \rceil \right). \tag{2.1.4}$$

Then if $N - m$ observations are drawn in the second stage the risk over all N observations will be less than a predetermined risk bound $w : R_N \leq w$

Proof.

We can re-express the risk on all N observations as $R_N = AE(\frac{\overline{X}_N}{\alpha} - \lambda)^2$ side of the inequality as

$$AE \left[\frac{m^2}{N^2} \left(\frac{\overline{X}_m}{\alpha} - \lambda \right)^2 + \frac{\lambda^2}{\alpha} \left(\frac{N - m}{N^2} \right) \right].$$

Recall $m \leq N$, so the ratio is $\frac{m}{N} \leq 1$, and $N \geq \frac{B\overline{X}_m^2}{\alpha^3 w}$.

Now,

$$AE \left[\frac{m^2}{N^2} \left(\frac{\overline{X}_m}{\alpha} - \lambda \right)^2 \right] \leq AE \left[\frac{m}{N} \left(\frac{\overline{X}_m}{\alpha} - \lambda \right)^2 \right]$$

$$\leq \frac{m\alpha^3 w}{B} AE \left(\frac{1}{\alpha^2} - \frac{2\lambda}{\alpha \overline{X}_m} + \frac{\lambda^2}{\overline{X}_m^2} \right)$$

Also,

$$\frac{\lambda^2}{\alpha} AE(\frac{N - m}{N}) = \frac{\lambda^2}{\alpha} AE(\frac{1}{N}(1 - \frac{m}{N})) \leq \frac{\lambda^2}{\alpha} AE(\frac{1}{N}) \leq \frac{\alpha^2 w}{B} AE_\lambda \frac{\lambda^2}{\overline{X}_m^2}$$

Thus, using the two inequalities above with the reexpression fact we have

$$AE \left(\frac{\overline{X}_m}{\alpha} - \lambda \right)^2 \leq \frac{m\alpha^2 w}{B} AE \left(\frac{m}{\alpha} - \frac{2m\lambda}{\overline{X}_m} + \frac{m\alpha\lambda^2}{\overline{X}_m^2} + \frac{\lambda^2}{\overline{X}_m^2} \right)$$

Using the fact that $\sum_{i=1}^{m} X_i \sim \text{Gamma}(m\alpha, \lambda)$, it is easily seen that $2\lambda^{-1}\overline{X}_m$ will be distributed χ^2 with $2m\alpha$ degrees of freedom. It can be verified that expectation will be $\frac{\Gamma(m\alpha-k)}{2^k\Gamma(m\alpha)}$. We obtain the equation:

$$R_N = AE\left(\frac{\overline{X}_N}{\alpha} - \lambda\right)^2 \leq \frac{A\alpha^2 w}{B}\left[\frac{m}{\alpha} - \frac{2m^2\Gamma(m\alpha-1)}{\Gamma(m\alpha)} + \frac{(m^3\alpha+m^2)[\Gamma(m\alpha-2)]}{\Gamma(m\alpha)}\right].$$

So to ensure the expected loss is less than our risk bound w, we set the righthand of the inequality to equal w then solve for B accordingly and obtain equation (2.1.5).

2.2 Improving the Terminal Sample Size

In the prior section, we found results that certainly achieved the goal of having the risk within the risk bound. An alternative to the asymptotic sampling that ensured the risk is within a bound was found. Remember that is only part of the goal; the goal is to sample the fewest number of observations that achieves the bounded risk goal. It is important to investigate the relationship between the terminal sample size and the optimal sample size. Exploring the relationship of the N and n^* is the first step in seeing if N needs to be reduced. If $m < n^*$, then

$$E[N/n^*] = \left[\frac{m}{\alpha} - \frac{2m^2\Gamma(m\alpha-1)}{\Gamma(m\alpha)} + \frac{(m^3\alpha+m^2)[\Gamma(m\alpha-2)]}{\Gamma(m\alpha)}\right]\left[1+(m\alpha)^{-1}\right].$$

This means on average, the terminal sample size will be larger than the optimal for any value of $m\alpha \geq 3$. This is what is meant by the procedure being exact. Clearly, the terminal sample size N is a biased estimator of n^*. Thus the asymptotic performance will be examined in similar fashion to equation 1.3.2 and 1.3.3. Notice also that in these equations the terminal sample size is a function of m, but does not necessarily consist of the m observations. Our procedure is a genuine two-stage sampling procedure so it will consist of m initial observations along with additional observations. Since that is the case, we cannot evaluate the asymptotic performance by simply looking at $m \to \infty$, because $N/n^* \to \infty$ as well.

However, it can be evaluated in the following manner $w \to 0$ as $m \to \infty$ and $E[N/n^*] < \infty$

$$\lim\, E[N/n^*] = \lim E[\frac{B\bar{X}_m^2}{\alpha^3 w}\frac{\alpha w}{A\lambda^2}]$$

$$= (A\alpha^2\lambda^2)^{-1} \lim E[B\bar{X}_m^2]$$

$$= (A\alpha^2\lambda^2)^{-1} \lim \left(A\alpha^2 \left[\frac{m}{\alpha} - \frac{2m^2}{m\alpha - 1} + \frac{m^3\alpha + m^2}{(m\alpha-1)(m\alpha-2)} \right] \right) E[\bar{X}_m^2]$$

$$= \lambda^{-2} \lim \left(\frac{m}{\alpha} - \frac{2m^2}{m\alpha - 1} + \frac{m^3\alpha + m^2}{(m\alpha-1)(m\alpha-2)} \right) E[\bar{X}_m^2]$$

$$= \lambda^{-2} \lim m \left(\frac{(m\alpha-1)(m\alpha-2) - 2m\alpha(m\alpha-2) + m^2\alpha^2 + m\alpha}{\alpha(m\alpha-1)(m\alpha-2)} \right) E[\bar{X}_m^2]$$

$$= \lambda^{-2} \lim m \left(\frac{m^2\alpha^2 - 3m\alpha + 2 - 2m^2\alpha^2 + 4m\alpha + m^2\alpha^2 + m\alpha}{\alpha(m\alpha-1)(m\alpha-2)} \right) E[\bar{X}_m^2]$$

$$= \lambda^{-2} \lim m \left(\frac{2m\alpha + 2}{\alpha(m\alpha-1)(m\alpha-2)} \right) E[\bar{X}_m^2]$$

Now $E[\bar{X}_m^2] = \alpha\lambda^2/m + (\alpha\lambda)^2$, which implies $E[\bar{X}_m^2\lambda^{-2}] = \alpha/m + \alpha^2$.

$$\lim E[N/n^*] = \alpha \lim \left(\frac{2m\alpha + 2}{\alpha(m\alpha-1)(m\alpha-2)} \right) + \alpha^2 \lim \left(\frac{m(2m\alpha + 2)}{\alpha(m\alpha-1)(m\alpha-2)} \right).$$

Thus,

$$\lim E[N/n^*] = 2. \tag{2.3.1}$$

In addition to studying the mean between the ratios of the terminal sample size to optimal sample size, the same should be done with the variance. It is important to consider the amount of variation that will occur for the best possible scenario.

$$\lim Var[N/n^*] = \lim Var[\frac{B\bar{X}_m^2}{\alpha^3 w}\frac{\alpha w}{A\lambda^2}]$$

$$= \lim(A^{-2}\lambda^{-4}\alpha^{-4}) Var[B^2\bar{X}_m^2]$$

Now $Var[\bar{X}_m^2] = E(\bar{X}_m^4) - (E\bar{X}_m^2)^2$. Recall $\bar{X}_m^2 \sim Gamma(m\alpha, \lambda/m)$. Using the moment generating function (1.4.4), one can easily find that

$$\lim Var(\bar{X}_m^2) = \alpha^2\lambda^4/m^2 - 2\alpha^3\lambda^4/m.$$

which means,

$$\alpha^2 \lim \left(\frac{4m^2\alpha^2 + 8m\alpha + 4}{\alpha^2(m\alpha-1)^2(m\alpha-2)^2} \right) - 2\alpha^3 \lim \left(\frac{m(4m^2\alpha^2 + 8m\alpha + 4)}{\alpha^2(m\alpha-1)^2(m\alpha-2)^2} \right).$$

Thus,

$$\lim Var[N/n^*] = 0. \qquad (2.3.2)$$

The limiting mean of the ratio between terminal sample size and optimal sample size is two. Also, the limiting variance of the ratio between the terminal sample size and optimal sample size is zero. Therefore, it can be concluded that the terminal sample size becomes twice that of the optimal sample size. Practically this means in the best possible scenario the terminal sample size will still be nearly twice the optimal on average. This might be the price of having a “genuine” two-stage sampling procedure that uses exact methodology. However, it is desired to reduce N so that its corresponding risk is just within the bound. It is important to recall the bound coefficient B is proportional to our terminal sample size N. With this in mind, it would be an improvement if an alternative bound coefficient B could be found. In order to truly see improved results, we must find a way to reduce the bound coefficient significantly. Instead the next section considers a more practical application. It will continue to use the B in result (2.1.5) and discuss reducing the sample size empirically to find a better bound coefficient through simulations.

This section studies R_N as a function of λ in order to determine where the maximal risk occurs. This is done because the risk function can be altered by some constant and yielding a new bound coefficient B_{new} that should reduce sample size and continue to bound the risk of the mean. Table A.3 gives values of the bound coefficient when $A = 1$. We see as both α and m increase B nears two. This gives some information about an appropriate sample size, but a better value for bound coefficient can be found that will give smaller values for the terminal sample size that will be closer to optimal sample size. Remember this is a generalization of the Exponential case. Zacks and Muhkopadhyay (2006) were faced with the exact same problem. In their article, the authors decide they can reduce B by investigating the distribution of the risk under their sampling procedure. This is done by identifying what value for λ gives the maximal risk, and afterwards empirically increasing B so that the maximal risk is just within the bound. Once that was done the new empirical

B was formed as a ratio of their prior B. The following are their results of these simulations:

$$B_{new} = 0.565B. \tag{2.3.3}$$

Since as α increases B decreases, simply choosing $0.565B$ will be significant improvement for the Gamma case for any $\alpha > 1$. No further work needs to be done. However, the fact that as α increases B decreases we choose to further our research and develop a new bound coefficient as a function of the old B and α.

The risk function of our two-stage sampling procedure was investigated as a function of the scale, in order to approximate the λ where the maximal risk occur. Clearly, larger λ values result in larger risks if N were to remain constant. However, larger λ values tend to result in larger values for N, which reduce the risk. So the maximal risk under this two-stage sampling procedure is not necessarily an infinite entity. In fact, through simulations the maximal risks most commonly occurred between five and six. Once this was done, B was identified for each α, then empirically reduced so that the risk is just within the bound w. The new B is the ratio of the empirical B found to the B as a result of mathematical theory given in (2.1.5). In Figure B.1, we can see a scatter plot of these ratios and α.

For each value of α there was a corresponding ratio. For example, $\alpha = 1$ would correspond to 0.560. Looking at the scatter plot, there appears to be a negatively exponential relationship between α and what the appropriate ratio should be, leveling off around $\alpha = 20$.

A regression is performed to exploit this relationship. This is done only to find coefficient of log (α). For this problem, we are not trying to fit the curve, but have a curve that gently sits above each of the points. In order to do this, the same regression is used but the slope needs to be altered. As stated earlier, 0.565 suffices for all $\alpha > 1$, this will be used to find the intercept. With this we find that:

$$\begin{aligned} B_{new} &= [-0.031\log(\alpha) + 0.597]B \quad , \quad \alpha < 20 \\ B_{new} &= 0.505B \qquad\qquad\qquad\qquad , \quad \alpha \geq 20 \end{aligned} \tag{2.3.4}$$

This will give smaller values for B see Table A.4.

2.2.1 Performance Properties of New Estimation Procedure

A genuine two-stage procedure for sampling data to an assigned accuracy only assuming the Gamma distribution was found in Theorem 2.1. This section showed that ultimately this procedure would continue to sample nearly twice as many observations than needed. A more practical solution of how to select a smaller number of observations was also considered in this section. Unlike section 2.1, a mathematically rigorous proof was not provided. However, sufficient analysis was performed to substantiate the belief that the risk will always be within the risk bound w and that a resulting reduction in terminal sample size of nearly half the observations will be selected. In fact if $\alpha = 1, w \to 0$ as $m \to \infty$ and $E[N/n^*] < \infty$ then

$$\lim E[N_{new}/n^*] = 1.13,$$

and if $\alpha \geq 20$, then

$$\lim E[N_{new}/n^*] = 1.01.$$

This means that instead of sampling nearly twice as many observations, the improved two-stage sampling procedure will sample between approximately 1.01 and 1.13 depending on the value of α and the initial sample size. This is seen in Table A.4. Not only does this improved result give reliable estimators for λ but the terminal sample size nears the optimal sample size.

2.3 Computer Simulations

In the previous section, the mathematical theory was provided to ensure the risk stays with the predetermined bound. To verify the results in the previous section a simulation study was conducted using R software. In the simulation, differing values for optimal sample size n^* were chosen: 25, 50, 100, 500. We fix $\lambda = 5$ since this result is not dependent upon

λ and vary $\alpha = \{0.5, 1, 2, 5, 10\}$. A is a constant expression, we choose it to be 2; 10,000 replications were used for each case. The quantity $\overline{N}$ is an estimate for the expected value of N and $\overline{r}$ is an estimate of the risk with the original terminal sample size. This simulation was repeated with the improved terminal sample size N_{new}. Our desired result is to see $\overline{r}$ fall beneath w and to see $\overline{r}_{new}$ be just below w and $\overline{N}_{new}$ to be above n^*. Also, since this is a generalization of Mukhopadhyay's research we would like to see the same results with $\alpha = 1$. As such, our results for $\alpha = 1$ and $\lambda = 5$ should resemble Mukhopadhyay and Pepe's (2006) results. Figure B.2, B.3, B.4, and B.5 give visual representation of the estimated risks compared to their risk bounds as shape varies 0.5, 1.0, 2.0, 5.0, 10.0 and n^* is fixed to 25. Figure B.4 and B.5 display the same for the improved results. Notice, the estimated risks fall within the risk bound. With the improved results the estimated risks are closer to the risk bound. Further detail is given in Table A.1 and Table A.2.

See Table A.1 and Table A.2, we note the following:

1) In Table A.1, for the case $\alpha = 1$ and $\lambda = 5$, our mean value for N and mean value for r are nearly identical as those given by Mukhapadyay and Pepe. The B given in this article is exactly that of their B. This is to be expected because this is a generalization of their result.

2) In Table A.1, the mean value for N nears twice that of n^*. Note N is a function of α and m, so as both variables get larger N gets closer to $2n^*$. We can see this if we look across rows and down columns.

3) In Table A.1, the mean value for r is always nearly half of our predetermined risk w. This follows since the expected risk is inversely proportional to the number of observations drawn.

4) Naturally as the initial sample size m increased, we obtained more information about the sample with which to make our decision and consequently we obtained better values for

r and N.

5) From Table A.1 and A.2, we observe that at no point in this simulation do the estimates for the risk nor the improved risk ever exceed the risk bound w.

6) In Table A.2, we observe that the newer results risks are much closer to the bound, and the newer sample sizes have been reduced on average by 43%.

7) In Table A.2, we observe that the average terminal sample size $\overline{N}_{new}$ is still larger than n^*. It is my conjecture, that it will be unable to improve N_{new} any further. Our average risks are just within the bounds, which is our goal. Reducing the sample size any further might result in having the average risk eclipse the risk bound.

2.4 Shape Unknown and Scale Unknown

In this section, we provide a solution to the question of how many samples should be selected if both parameters are unknown. If variables are both unknown, finding bounds become more difficult. The goal remains the same, to find an appropriate sample size N that will make the associated *risk* function less than or equal to a predetermined risk $w > 0$. Recall if $X \sim Gamma(\alpha, \lambda)$ then the mean is $\alpha\lambda$ and the estimator for the mean is $\overline{X}$. The goal is to find N such that:

$$AE(\overline{X} - \alpha\lambda)^2 < w,$$

This problem cannot be solved mathematically as it was done in section 2.1. The proof in said section requires the chi-square transformation that enables us to find the expectation without knowing the scale parameter. Without knowing the shape parameter that transformation cannot be used. As stated earlier, studying the shape parameter known case lays the ground work for times that the shape parameter is unknown.

Like many other statistical procedures, in place of the parameter an estimate will be used which will yield an estimate for bound coefficient and an estimate for the terminal sample size. The first stage is to collect m initial observations and to find an estimate of both α and the mean. The second stage collects $\hat{N}$, where $\hat{N}$ is as follows

$$\hat{N} = \max(m, \lceil \frac{B\overline{X}_m^2}{\hat{\alpha}^3 w} \rceil). \tag{2.4.1}$$

The law of large numbers guarantees that as the sample size approaches infinity the estimate will approach its true parameter value. This means that with a relatively large initial sample size, the result should work nearly as well as in section 2.1. In section 2.5, we address how the initial sample size should be selected.

At this point, the next step is to evaluate the performance of terminal sample size. The method of moments estimator is used for the shape ($\hat{\alpha} = (\bar{X}_m/S_m)^2$). As we mentioned in section 1.4, there is no closed form maximum likelihood estimator for α. The sample mean will be used as an estimate for the true mean. The simulations are set up the same as before. In the simulation both parameters are known and differing values for optimal sample size n^* were chosen: 25, 50, 100, 500; 10,000 replications was ran for each simulation. The quantity $\overline{N}$ is an estimate for the expected value of $\hat{N}$ and $\overline{r}$ is an estimate of the risk. We would like see $\overline{r}$ be just below w and $\overline{N}$ to be above n^*. For the first result, we fix $\lambda = 5$ and vary $\alpha = \{1, 2, 5, 10\}$. Also, since A is a constant expression we choose it to be 2. This way, our results for $\alpha = 1$ and $\lambda = 5$ should resemble Mukhopadhyay and Pepe's (2006) results. The simulations are given in Table A.5 and Table A.6. Figures B.2 and B.3 show a visual representation of how the estimated risks compared to the risk bounds, as α varies 1, 2, 5, 10. Notice how it consistently falls below the risk bound.

1) As we might imagine when m is small the numbers differ greatly, because determining N is heavily dependent on α. The $m = 10$ observations were inconsequential and not even

worth recording. The larger the initial sample size, the better estimate of α we will obtain.

2) With a non-constant α present, there is more variance in the estimates for N. Note: $\overline{N}$ is larger than the α known counterpart, yet the risk is higher. This is due to skewed unknown distribution of N. Even though there is more variability among the statistic, one should not expect to see the bound exceeded as long as cautionary measures are taken.

3) The estimates for the still risk fall below the risk bounds most of the time. There are cautionary factors when the initial sample size is too small, as well as cautionary factors when the risk bound w is very small.

2.4.1 Robustness Considerations

No additional simulations are necessary to conclude that this method is not robust to the α known assumption. As initial sample size gets large the bound coefficient nears A. However, since terminal sample size has a cubic α term in its denominator, being off by the smallest margin adversely impacts its value greatly. This is shown in an example in section 2.6. This was seen in the $m = 10$ case where the differences were so severe they were not worth reporting. When both parameters are unknown, one should take a moderate size initial sample.

2.5 Determining Initial Sample Size

One might note that bound coefficient B_{new} is a function of the initial sample m as well as the shape parameter α. Determining the initial sample size is very important in achieving the goal of sampling the optimal amount. The procedure as proposed in Theorem 2.1 does not specify m. This section gives a method of selecting the initial sample as long as the user has a vague idea of the parameters. The problem with blindly selecting m initial observations is as follows: if m is chosen too small then the terminal sample size N will be large and if m is chosen too large then the terminal sample size N will then equal m and consequently be too large. It is desired to select m initial observations that minimize the terminal sample size N. Before details are laid out, it is important to note that this is a practical application and is best if discretionary measures are used.

The sampling procedure given in section 2.1, is a genuine two-stage sampling procedure. This means no additional statistical information is given prior to the first stage and no data can be used to determine the initial sample m. This does not mean that there is no general knowledge about the population being sampled. It is possible that there is mathematical theory or historical evidence to determine what the parameters might be. Those values should yield a decent value for m that will not inflate the terminal sample size. Notice that the optimal sample size given in (2.1.3) is a function of the parameters α and λ.

STAGE 1:

$$m = \lceil \frac{A\lambda_0^2}{\alpha w} \rceil$$

STAGE 2:

$$N = \max\left(m, \lceil \frac{B_{new}\bar{X}_m^2}{\alpha^3 w} \rceil\right)$$

In all the previous simulations the initial sample size was preset $m = \{10, 20, 30\}$. As such we saw how the terminal sample size improved as the initial sample size increased. However, this section points out there is a risk associated with a large initial sample size.

We will see if this leads to a reduction in the terminal sample size. In the simulations, shape and scale is equal to two and five respectively, A is chosen as two like before, w is chosen to be 0.500, 0.250, and 0.125. To compare this idea to the one prior, the values for m are 5, 10, 50, 100, and 500, and then that is compared to m values if the hypothesized value is within 25% of the true scale parameter, which is 3.75 and 6.25 respectively.

Table A.7 shows that if the hypothesized value is within 25% of the true scale, then the resultant terminal size in each example is smaller than when m is chosen too small, i.e. $m = 5$ and $m = 10$ and when it is chosen too large $m = 500$. We caution the user to use discretionary measures. If the researcher is not confident that their hypothesized value is even close to the unknown parameter then it will be better to sample a decent size initial sample size as they see fit.

2.6 Example in Understanding Precipitation Rates in Regional Climate Models

Maureil et al (2007), Gutowski et al (2008), and Groisman et al (1999) all state that precipitation rate intensities can be modeled under a Gamma distribution. In this example we will model the precipitation rate intensity with the Gamma distribution. Furthermore, we will estimate rainfall in the West Point, GA, United States using the sequential estimation proposed earlier sections. Their are two purposes of this example, the first illustrate the bounded risk sampling procedure according to their theory regarding the shape, the second is to see how robust their shape assumption is by modeling the data with the with varying shape parameters. This is purely an example of how to use this bounded risk estimation procedure; there is no effort to solve any of their climatology problems.

Gutowski et al. (2008) notes that total precipitation in a bin (referring to histograms) may increase under the warming scenario, but its relative contribution to total precipitation may decrease. A positive change in bins of normalized distribution, not only have greater precipitation in the scenario climate, but they contribute relatively larger amounts to the total. Groisman et al (1999) analysis reveals increases in extreme precipitation provide evidence for statistically significant increases in precipitation in the United States. These climate models have projected increase in global precipitation, which is believed to be due to global warming stemming from increases in greenhouse gases.

Gutowski explains the theoretical model of intensity of daily precipitation.

$$p(x) = p_o x^{\alpha-1} exp(-x/\lambda), \tag{2.6.1}$$

where, p_o and λ are parameters of the distribution and a restriction $\alpha \geq 1$. The total precipitation during a period described by

$$P = \int_0^\infty p_o x^{\alpha-1} exp(-x/\lambda) dx. \tag{2.6.2}$$

and the total number rain days is

$$N = \lim_{\epsilon \to 0} \left[\int_0^\infty p_o \frac{x^{\alpha-1}}{x} exp(-x/\lambda) dx \right]. \tag{2.6.3}$$

Normalizing equation (2.6.1) by dividing the total precipitation yields the Gamma distribution, see (1.4.1)

$$p(x) = \frac{x^{\alpha-1} exp(-x/\lambda)}{\lambda^\alpha \Gamma(\alpha)}. \tag{2.6.4}$$

Gutowski further believes that the shape parameter should be two in the regions they study. They state it is not a requirement for $\alpha = 2$, however the case $\alpha = 1$ poses problems for computing the number of rain days (2.6.2) and is not physically realizable in the present context. This is an example of when shape is known and scale is unknown.

The data used in this example was collected from the United States Historical Climatology Network. Here, we will look at one city in the southeastern United States; West Point, GA during the warm season, which is defined by Gutowski as (April - September). Forty initial observations were collected from the years 2001-2005. Based on our value for N we will make the decision to collect more observations if necessary. Assume that each of the five cities have equivalent distributions since we are modeling the region.

Let $A = 2.5$ and $w = 0.0025$.

Below, we can see the table of the data that was collected.

Precipitation of West Point, GA. A=2.5 w=0.0025 and α=2

Pilot Data			
m=40	$B = 5.258$	$B = 2.655$	
0.09, 0.55,	0.73, 0.05,	0.05, 0.01,	0.97, 0.23
0.54, 1.75,	0.51, 0.20,	1.15, 0.60	1.39, 0.32
0.66, 0.01,	0.35, 0.19,	0.61, 0.29	0.20, 3.30
0.49, 0.10,	0.47, 0.57,	2.00, 0.23,	0.61, 0.18
0.19, 1.00,	0.30, 0.07,	0.35, 0.62,	1.20, 0.86
$\overline{X}_m$=0.558	$\frac{B\overline{X}_m^2}{\alpha^3 w} = 41.338$	$\Rightarrow N = 42$	
New Data			
$N - m = 2$			
0.12, 0.02			
$\hat{\lambda} = 0.287$			

This process is repeated assuming the shape is 1.75 and again when the shape is 2.25. This is done to see how varying the shape will affect the total sample size and the mean of the entire sample. Below, we can see the table of the data that was collected.

Precipitation of West Point, GA. A=2.5 w=0.0025 and α=1.75

Pilot Data			
m=40	$B = 5.296$	$B = 2.674$	
0.09, 0.55,	0.73, 0.05,	0.05, 0.01,	0.97, 0.23
0.54, 1.75,	0.51, 0.20,	1.15, 0.60	1.39, 0.32
0.66, 0.01,	0.35, 0.19,	0.61, 0.29	0.20, 3.30
0.49, 0.10,	0.47, 0.57,	2.00, 0.23,	0.61, 0.18
0.19, 1.00,	0.30, 0.07,	0.35, 0.62,	1.20, 0.86
$\overline{X}_m$=0.558	$\frac{B\overline{X}_m^2}{\alpha^3 w} = 62.154$	$\Rightarrow N = 63$	
New Data			
$N - m = 64$			
0.12, 0.02,	0.25, 0.08,	1.43, 0.12,	0.04, 0.03
0.04, 0.26,	0.23, 0.76,	0.04, 0.22,	0.70, 1.30
1.30, 0.62,	0.55, 0.02,	1.10, 0.60,	0.07
$\hat{\lambda} = 0.307$			

This is done again, when $\alpha = 2.25$

Precipitation of West Point, GA. A=2.5 w=0.0025 and α=2.25

Pilot Data			
m=40	$B = 5.228$	$B = 2.640$	
0.09, 0.55,	0.73, 0.05,	0.05, 0.01,	0.97, 0.23
0.54, 1.75,	0.51, 0.20,	1.15, 0.60	1.39, 0.32
0.66, 0.01,	0.35, 0.19,	0.61, 0.29	0.20, 3.30
0.49, 0.10,	0.47, 0.57,	2.00, 0.23,	0.61, 0.18
0.19, 1.00,	0.30, 0.07,	0.35, 0.62,	1.20, 0.86
$\overline{X}_m$=0.558	$\frac{B\overline{X}_m^2}{\alpha^3 w} = 28.870$	$\Rightarrow N = 40$	
$\hat{\lambda} = 0.266$			

This procedure was done varying $\alpha = \{1.75, 2.00, 2.25\}$. Notice, that the values for B remained close at 2.674, 2.655, and 2.640 respectively. Even though that is the case, the values for $\lceil B\overline{X}_m^2\alpha^{-3}w^{-1}\rceil$ varied much more with 29, 42, and 63 respectively. So varying α only 25 tenths can lead to a large difference in the final sample size. There was a total of 246 raindays recorded at this station. The mean over all 246 observations were 0.540 and the means for the sample 0.599, 0.570, and 0.537. This is just an example to show that the two-stage sampling procedure will lead to a reliable estimate and to see how adjusting the shape parameter affects the final sample size. We sampled a total of 64 observations and is indeed within the risk bound specified earlier. It should be noted that we only looked at one city from the years 2001-2005. In fact, the United States Historical Climatology Network has 1,062 stations across the nation with some dating back before 1900. There is a wealth of information to develop a wide variety of climate models. There is a plethora of information and all of the data need not be used to develop a reliable estimate for precipitation intensity.

2.7 Discussion

It is well known that bounding the risk with a fixed-sample size is impossible. This is the reason a two-stage sequential estimation procedure was implemented. There is prior research involving genuine two-stage exact methods for a Normal and Exponential population, but no research in this area for a Gamma population.

We mathematically determined a sample size that will always ensure the risk is within a predetermined risk bound when the shape parameter is known, and an estimate for that sample size when the shape parameter is unknown. The consequence of a two-stage exact method was the end result of sampling more than twice as many observations as need be. The function of R_N through simulations is to aid us in arriving at a more practical solution.

Finally, this procedure was illustrated on precipitation of West Point, GA in the summer months of 2001-2005.

Chapter 3
Fixed Width Confidence Interval Estimation

The focus of the dissertation is developing reliable estimators using exact evaluation criteria. The criterion used in this chapter is having the interval estimator less than or equal to some predetermined width. Whereas, bounding the risk certainly gives some indication of the reliability of the estimator, an interval estimator will give more interpretable results.

Fixed-width confidence intervals are prevalent in sequential estimation. There are a number of articles that restrict the width of the mean for different distributions. There is, however, not a lot of research in this area on the Gamma distribution. Chow and Robbins (1965) develop a two-stage sampling for a general distribution $f(x)$, but their procedure uses Normal approximation. Govindarajulu (1995) developed a sequential estimator for the mean of an Exponentially distributed population. This result, though more specific to the exponential distribution still uses Normal approximation. This chapter uses pre-assigned risk to answer the question of how to restrict the interval estimator. In many ways these two topics are related, as discussed by Stein for the Normal distribution. Before our interval estimator is proposed, we review confidence intervals for the Gamma distribution.

3.1 Confidence Intervals for Gamma Distribution

Confidence intervals are one of the fundamental aspects of statistical inference. In this section, we will review former interval estimators for the Gamma distribution. We should mention significant research in the Gamma distribution is performed with shape known. We have already discussed why the shape known case is studied.

The following example comes from Casella and Berger (2002). The example these authors give pivot the statistic $2\sum_{i=1}^{n} X_i/\lambda$. Denote g^*_q as the q^{th} quantile of a Gamma distribution with shape $n\alpha$ and scale λ/n. This example involves inverting a statistic $P(g^*_{a/2} < \bar{X}_n < g^*_{1-a/2}) = 1-a$, and a is its significance level. The estimator of λ used is $\overline{X}_n/\alpha$. If $X_1, .., X_n$ are Gamma i.i.d. variables with shape α and scale λ, then we know $\bar{X}_n$ will be Gamma Distributed with shape $n\alpha$ and scale λ/n. We can multiply $\bar{X}_n$ by $2/\lambda$, and obtain a new variable $Y \sim \chi^2_{2n\alpha}$. Denote c_q as the q^{th} quantile of a Chi-Square distribution with parameter $2n\alpha$. So,

$$\begin{aligned}
&P(g^*_{a/2} < \bar{X}_n < g^*_{1-a/2}) \\
&= P(g^*_{a/2}/\alpha < \bar{X}_n/\alpha < g^*_{1-a/2}/\alpha) \\
&= P(\tfrac{2g^*_{a/2}}{\alpha\lambda} < 2\bar{X}_n/\alpha\lambda < \tfrac{2g^*_{1-a/2}}{\alpha\lambda}) = 1-a \\
&= P(\tfrac{2g_{a/2}}{\lambda} < 2\bar{X}_n/\lambda < \tfrac{2g_{1-a/2}}{\lambda}) = 1-a \\
&= P(c_{a/2} < 2\bar{X}_n/\lambda < c_{1-a/2}) = 1-a
\end{aligned}$$

Rearranging the equations, the $1-a$ interval estimator can be obtained

$$C_X = \{\lambda | \frac{2\bar{X}_n}{c_{1-a/2}} < \lambda < \frac{2\bar{X}_n}{c_{a/2}}\}. \tag{3.1.1}$$

Neither of the aforementioned exact confidence intervals can be restricted. The interval estimator presented in (3.1.1) is a multiple of $\bar{X}_n$, so it is dependent upon knowing all n observations.

Fixing interval estimators widths is prevalent in sequential estimation. A brief synopsis of fixed width confidence intervals, and formal definitions of confidence intervals were given in the first chapter. In this section, we explicitly define our goal in terms of the Gamma distribution. There are two components of a confidence interval: (1) the interval width and (2) the coverage probability. The interval width is the range from the lower bound $L(X)$ to the upper bound $U(X)$. The coverage probability refers to the probability that the true parameter is covered in that interval. For a fixed sample size, the two are inversely proportional to one another. There are merits to both components. A low coverage probability corresponds to high chances of the experimenter making an error; however, a large interval width makes the interval estimator uninformative. The goal is to find the sample size that will ensure a high coverage probability and a narrow interval width.

Notice how results (1.5.4), (1.5.5), and (1.5.6) all developed terminal sample sizes for fixed-width confidence intervals either by assuming the population is Normal or asymptotically will become Normal. This is due to the fact that these confidence intervals are traditionally of the form given in (1.5.3). Since our procedure is only assuming Gamma populations and it is known that it will be asymmetric; we emphasize our interval estimator will not be of that same form.

3.2 Two-Stage Fixed-Width Confidence Interval for Scale

In this section, we propose a two-stage sampling procedure that assumes that the observations come from a Gamma population. As mentioned earlier there is little research in this area and there is no research that uses bounded risk to arrive at an interval estimator for observations that are assumed to be Gamma. This procedure uses risk bounds developed in chapter 2 to develop an upper bound for λ.

For observations $X_1, X_2 ...$ i.i.d. Gamma distributed variables with shape α and scale λ. For a predetermined confidence interval width d, the goal is to estimate the mean with $1 - a$ coverage probability, less than or equal to the width d. That is, $P(\alpha\lambda \in C_X) \geq 1 - a$. As in earlier sections, we have been assuming shape is known. If shape is known then our goal becomes:

$$(1) \quad P(\lambda \in C_X) \geq 1 - a$$

$$(2) \quad C_X \leq d$$

Note that in the introduction it was listed as $C_X \leq 2d$. The reason for this is due to the fact that in most of the prior work in this area the observations were assumed to come from symmetric distributions. When that is the case, the mean is no longer a factor. For a symmetric distribution $C_X = \{\mu | \bar{X}_n - d < \mu < \bar{X}_n + d\}$, the width is $2d$ and completely independent of the sample mean. As long as there is knowledge of the variance, knowledge of mean is not required. This is a luxury that asymmetric distributions such as the Gamma do not have. Notice that the interval estimator (3.1.1) is a multiple of the sample mean. The distance from the sample mean to the respective upper and lower bounds will not be the same for the interval estimator given in this section.

The optimal sample size is the first n for which both criteria is achieved.

$$n^* = \min\{n \in \mathbb{N} | P(\lambda \in C_X) \geq 1 - a, C_X \leq d\}. \tag{3.1.5}$$

According to Ghosh (1991), ideally the terminal sample size for a fixed-width confidence interval should have the following properties:

(1) N is non decreasing in $d > 0$.

(2) N is finite with probability 1 for every $d > 0$.

(3) $N/n^* \to 1$ as $d \to 0$ in probability or a.s.

(4) $E(N)/n^* \to 1$ as $d \to 0$.

(5) $\lim_{d\to 0} P(\lambda \in C_X) = 1 - a$

In the following section, we will show that under certain conditions these properties hold.

Theorem 3.1 *For significance level a and predetermined width d, if $X_1, ..., X_m$ i.i.d. Gamma (α, λ) initial observations are drawn ($m\alpha \geq 3$). If N is defined in 2.1.4 and g_q be the qth quantile of the Gamma$(n\alpha, 1/n)$ distribution*

$$M = \min\{n \geq N \in \mathbb{N} | \sqrt{\frac{Nw\alpha}{A}}[\frac{g_{1-a/2} - g_{a/2}}{\alpha}]\} \tag{3.1.6}$$

Then if $C_X = \{\lambda | \bar{X}_M/\alpha - \sqrt{\frac{Nw\alpha}{A}}[g_{1-a/2}/\alpha - 1] < \lambda < \bar{X}_M/\alpha - \sqrt{\frac{Nw\alpha}{A}}[g_{a/2}/\alpha - 1] = d\}$ Then

(1) $P(\lambda \in C_X) \geq 1 - a$

(2) $C_X \leq d$

Proof.

Theorem 2.1 ensures that $\dfrac{A\lambda^2}{\alpha N} < w$. This means,

$$\lambda < \sqrt{\frac{Nw\alpha}{A}}.$$

Denote g^*_q be the qth quantile of the $Gamma(n\alpha, \lambda/n)$. The sampling distribution of $\bar{X}_n \sim Gamma(n\alpha, \lambda/n)$,

$P(g^*_{a/2} < \bar{X} < g^*_{1-a/2}) = 1 - a.$

Let g_q be the qth quantile of the $Gamma(n\alpha, 1/n)$ distribution. Due to the scale property (1.4.2),

$P(\lambda g_{a/2} < \bar{X}_n < \lambda g_{1-a/2}) = 1 - a$

$P(\lambda g_{a/2}/\alpha < \bar{X}_n/\alpha < \lambda g_{1-a/2}/\alpha) = 1 - a$

$P(\lambda g_{a/2}/\alpha - \lambda < \bar{X}_n/\alpha - \lambda < \lambda g_{1-a/2}/\alpha - \lambda) = 1 - a$

$P(-\bar{X}_n/\alpha + \lambda[g_{a/2}/\alpha - 1] < -\lambda < -\bar{X}_n/\alpha + \lambda[g_{1-a/2}/\alpha - 1]) = 1 - a$

$P(\bar{X}_n/\alpha - \lambda[g_{1-a/2}/\alpha - 1] < \lambda < \bar{X}_n/\alpha - \lambda[g_{a/2}/\alpha - 1]) = 1 - a$

$P(\bar{X}_n/\alpha - \sqrt{\frac{Nw\alpha}{A}}[g_{1-a/2}/\alpha - 1] < \lambda < \bar{X}_n/\alpha - \sqrt{\frac{Nw\alpha}{A}}[g_{a/2}/\alpha - 1]) \geq 1 - a$

The width of the confidence of this confidence interval is $\sqrt{\frac{Nw\alpha}{A}}[\frac{g_{1-a/2}-g_{a/2}}{\alpha}]$. This is set to width d, and M is found accordingly.

$\sqrt{\frac{Nw\alpha}{A}}[\frac{g_{1-a/2}-g_{a/2}}{\alpha}] = d$

No close form solution exists, but the numeric solution yields the interval estimator given in (3.1.6).

3.3 Computer Simulations

To verify this sampling procedure gives accurate results, a number of simulations were performed. First to verify that the terminal sample size does increase as the predetermined bound d decreases and secondly to see how w affects the terminal sample size.

For the first simulation, the shape parameter was varied $\alpha = \{1, 2, 5, 10\}$, $w = \{5, 4, 3\}$, and $d = \{10, 5.0, 2.5, 1.0\}$. A was fixed at one. The value chosen for A should not be a large contributing factor to the final sample size M since the expectation of $\sqrt{Nw\alpha/A}$ is no longer a factor of A. However, one should use discretion since $M \geq N$ and N is dependent A, a large value for A might result in an inflated number for M. This simulation used 10,000 replications from a Gamma population with scale equal to five.

For the second simulation, the goal is to see if the percentage of times the scale parameter is within our interval and is greater than the $1-a$ coverage probability. The shape parameter was varied $\alpha = \{1, 2, 5, 10\}$, $w = \{5, 4, 3, 2, 1\}$, and λ fixed to be five. With 1,000 replications, we observe the percentage of times the parameter lies within the confidence interval. This was done for 80%, 85 %, 90%, and 95% coverage probabilities.

For the third simulation, it is desired to just see how the initial sample size m affects the percentage of times the scale parameter is covered. The risk bound was varied $w = \{5, 2, 1, 0.5, 0.2, 0.1\}$ and the initial sample size was varied $m = \{4, 6, 8, 10, 12\}$.

Notice the following:

1) Table A.8 shows that as the width bounds became smaller the sample size does increase.

2) Table A.9 shows that smaller risk bounds yield better initial estimates for the scale parameter. However, choosing w to be too small will inflate N which will consequently inflate M. Similarly, if m is chosen too large it will inflate N.

3) Table A.9 shows that the percentage that the parameter is within the confidence interval is always greater than the $1-a$ confidence level.

4) As always, initial sample size plays a factor in the estimate. Since the $E(N) \geq B\alpha^{-3}w^{-1}[\alpha^2\lambda^2 + \alpha\lambda^2/m] = B\lambda^2(\alpha w)^{-1} + B\lambda^2(m\alpha^2 w)^{-1}$, it was suspected that large m values will yield closer to the exact distribution. However, the risk bound w plays more of a factor than m does.

5) This is a numeric solution, so the researcher needs a maximum number of observations they are willing to sample in order to yield a solution. For these simulations our threshold

maximum was 50,000. This threshold maximum will affect the average M value. Simulated values for M might be biased above because of this reason.

6) Table A.10 shows that the initial sample size contributes to the estimated coverage probability. However, the risk bounds seem to contribute more than the initial sample size.

3.4 Asymptotic Performance

Much like sequential risk estimators, it is pivotal that the terminal sample size of this procedure be assessed. It will be shown that: (1) the ratio of expectations between the terminal sample size and the optimal sample size is greater than one, (2) how well this procedure will perform under certain conditions to see if the properties given in (3.1) will hold. Unfortunately there is no close form solution of the terminal sample size nor the optimal sample size. It is recorded that M will be the first sample size greater than N such that the equality (3.1.6) holds. Likewise, the optimal sample size is the integer n such that the probability is equal to $1-a$ and distance is equal to d. Recall that g_q is the qth quantile of a Gamma distribution with mean one and variance $(n\alpha)^{-1}$. By inverting the distribution of mean estimator, it can be found that the optimal sample size for the interval estimator is

$$n^* = \min\{n > 0|\lambda\alpha^{-1}(g_{1-a/2}(m) - g_{a/2}(m)) = d\}.$$

The ratio of expectations $E[M/n^*]$ becomes

$$E\left[\frac{\min\{n > N|\sqrt{\frac{Nw\alpha}{A}}\alpha^{-1}(g_{1-a/2}(M) - g_{a/2}(M)) = d\}}{\min\{n > 0|\lambda\alpha^{-1}(g_{1-a/2}(m) - g_{a/2}(m)) = d\}}\right].$$

Clearly if $n^* > N$, this reduces to

$$E\left[\lambda^{-1}\sqrt{\frac{Nw\alpha}{A}}\right].$$

Thus,

$$E[M/n^*] \leq \sqrt{1 + (m\alpha)^{-1}}.$$

It is obvious that as $d \to 0$, $M \to \infty$. If the terminal sample size increases the random variable $(G/\alpha - 1)$ should be examined asymptotically. It was mentioned earlier that $G \sim Gamma(M\alpha, 1/M)$. This means that the mean and variance of G/α will be one and $(M\alpha)^{-1}$ respectively. According to the Central Limit Theorem

$$\sqrt{M\alpha}(G/\alpha - 1) = Z \sim N(0, 1).$$

Also, the random variable $\bar{X}_M/\alpha \sim Gamma(M\alpha, \lambda/M\alpha)$ with the standard deviation (σ) equaling $\lambda/\sqrt{M\alpha}$. Let n_0^* be the optimal sample size of the risk estimator. If $m \to \infty$ and $w \to 0$ such that $E[N/n_0^*] < \infty$ and $d \to 0$.

$$1.01\lambda < \lim E\left[\sqrt{\frac{Nw\alpha}{A}}\right] < 1.13\lambda$$

Under those conditions, the upper bound of the proposed interval estimator

$$\bar{X}_m/\alpha - \sqrt{\frac{Nw\alpha}{A}}(g_{a/2}/\alpha - 1).$$

approximately becomes

$$\bar{X}_m/\alpha - \frac{\sigma}{M}(z_{a/2}).$$

Because the Normal distribution is symmetric it is the same as

$$\bar{X}_m/\alpha + \frac{\sigma}{M}(z_{1-a/2}).$$

Similarly, the lower bound will be the same lower bound as the Normal lower bound. The Normal distribution will have all the optimal properties. Thus, under those conditions asymptotically the proposed procedure will have all of the optimal properties. To summarize, the procedure was developed using exact methodology and will hold for any number of initial observations ($m\alpha > 3$). A small initial sample size may come with a price of sampling more observations than needed. For large m such that $m < n_0^*$ and $N < n^*$, the proposed confidence interval becomes approximately Normal and preserves many optimal properties. Both n_0^* and n^* are unknown quantities, so future research might entail finding a procedure that does not have to succumb to all of these exceptions

3.5 Example in Air Force Aeronautical Maintenance

In this section, we will show that these statistical estimation procedures can be used in real life situations. For multiple purposes, the United States Air Force needs to assess the readiness of the Air Force fleets. When a large number of planes are not operational the fleet has a low readiness, which might consequently put the United States nation at high risk, described by Rodrigues et al. (2000) and Morales et al. (2007).

Unspecified component time-to-failures are modeled with an Exponential distribution, Morales et al. (2007). In order to do this, researchers must conduct an experiment to collect data to see average lifespan of the component. There are three reasons a sequential framework is suited here: (1) the experiment might involve destruction of the component, (2) the time measured to failure as well as the time measured to repair is measured in days, and (3) to find the average service time. The compensation of each worker is an expense that must also be considered. This problem becomes two reliability estimation problems. The multistage layout allows them to reduce the price of conducting the experiment.

Though they mention modeling times as an Exponential distribution, they explicitly mention relaxing distributional assumptions from exponential to Gamma or Erlang.

It is desired to estimate operational availability of an air force plane that is defined by Kang (1998) as the ratio of estimated time operational over the estimated time operational and time not operational,

$$A_o = \frac{E(T_o)}{E(T_o) + E(T_{no})}.$$

This problem consists of four aspects: arrivals, service, finite population, and discipline. The planes single components fail over time and each component is believed to be Exponentially distributed. For simplicity, consider a single type of component for inventory. This corresponds to the time operational T_o. Service time for this example refers to the time that is required to repair a component. This corresponds to the time not operational T_{no}. The servers are the c repair crews. If one of the repair crews is idle, a broken part is repaired immediately; otherwise, it needs to wait in a queue until a crew gets idle. The repair times are assumed to be Exponentially distributed. This also contains the additional assumption that there is a finite population, which we can imagine in the context of planes there will not be an unlimited supply. In this example, the assumption is that the plane becomes operational immediately meaning the removal and installation times of broken/spare parts are negligible. The last assumption is that the queue is first in first out (FIFO) queue.

Dealing with government military real data is not readily available. Morales et al. (2007) constructed a sample of convenience with 250 repair and 250 life times by simulating from exponential distributions with rates 180 days/failure and 30 days/repair. N_o, r, and β are subjects specified in their article. Unlike, Morales' article our emphasis is not on the following goals:

Goal 1: Guarantee an average number of operative components at least equal to the required ready-to-fly r, $E(N_o) \geq r$, assuring that the mean number of operative planes, averaging over time, will be adequate for the required working fleet.

Goal 2: Assure a high probability of having at least r operative components available, $P(N_o \geq r) \geq \beta$, for a sufficiently large $\beta \in [0, 1]$. This establishes guarantees about the number of planes available at any time point.

Our goal is to use the information gained in this dissertation to estimate the repair and life times. A real life scenario will be created based off of this information to determine if fewer observations can be obtained to get a reliable estimate. Instead of simulating from rates of 180 days/failure and 30 days/repair our simulations will be from 200 days/failure and 25 days/repair. This information will be used to determine the initial sample size.

Also, so that this is more applicable to the Gamma environment we will assume that there is a spare present, making the shape parameter two.

The estimate for our operational availability is,

$$\hat{A_o} = \frac{\hat{T_o}}{\hat{T_o} + \hat{T_{no}}}.$$

Using 180 as hypothesized λ and 30 as hypothesized μ, we will determine if fewer samples can be used to make our estimates within 50 and 10 respectively. We allow $A = 1$ and $w = 150$. Remember these values are important for determining the sample size of the second stage N but should not affect the terminal sample size M.

$m = \lceil \frac{A\lambda_0^2}{\alpha w} \rceil = \lceil \frac{180^2}{2(150)} \rceil = 108.$

This means our initial sample will consist of 80 observations and will be used to find the second sample.

$N = \max\{m, \lceil \frac{B\bar{X}_m^2}{\alpha^3 w} \rceil\} = \max\{80, \lceil \frac{1.17(385.5)^2}{2^3 150} \rceil\} = 146.$

This means $\lambda < \sqrt{Nw\alpha} = 209.28$, this value is used in constructing the confidence interval.

$M_d = 134.4$ which implies $M = 146$. The estimate of λ over all 146 observations is 205.51. Finally the confidence interval is

$C_X = \{\lambda | \bar{X}_M/\alpha - \sqrt{\frac{Nw\alpha}{A}}[g_{1-a/2}/\alpha - 1] < \lambda < \bar{X}_M/\alpha - \sqrt{\frac{Nw\alpha}{A}}[g_{a/2}/\alpha - 1]\}$

$= \{\lambda | 205.51 - 209.28[2.235/2 - 1] < \lambda < 205.51 - 219.28[1.777/2 - 1]\}$

$= \{\lambda | 180.9 < \lambda < 228.8\}.$

Similarly, this is done with the service times. We allow $A = 1$ and $w = 10$. Remember these values are important for determining the sample size of the second stage N but should not affect the terminal sample size M.

$m = \lceil \frac{A\lambda_0^2}{\alpha w} \rceil = \lceil \frac{30^2}{2(10)} \rceil = 45$

This means our initial sample will consist of 45 observations and will be used to find the second sample.

$N = \max\{m, \lceil \frac{B\bar{X}_m^2}{\alpha^3 w} \rceil\} = \max\{45, \lceil \frac{1.17(45.39)^2}{2^3(10)} \rceil\} = 45.$

The value for $N = 45$, the rate μ has is bounded below $\mu < \sqrt{45(20)} = 30$, this value is used in constructing the confidence interval.

$M_d = 69.05$ which implies $M = 70$. Twenty-five additional observations need to be drawn.

$C_X = \{\lambda | \bar{X}_M/\alpha - \sqrt{\frac{Nw\alpha}{A}}[g_{1-a/2}/\alpha - 1] < \mu < \bar{X}_M/\alpha - \sqrt{\frac{Nw\alpha}{A}}[g_{a/2}/\alpha - 1]\}$
$= \{\lambda | 22.37 - 30.0[2.344/2 - 1] < \lambda < 22.37 - 30.0[1.682/2 - 1]\}$
$= \{\lambda | 17.21 < \lambda < 27.14\}$.

The length of the first confidence interval is 47.9, which is lower than our predetermined width of 50. The confidence interval length for the estimate is 9.93. For both component failure times and service times, the intervals contain the actual parameter. Finally, our estimate of the operational availability is:

$$\hat{A}_o = \frac{410.20}{410.20 + 44.74} = 0.902.$$

This actual statistic is distributed with a Beta distribution, and actual restrictions can be left for future research. We can also find the long run fraction of time that the queue is empty. In this particular example, an empty queue would mean that there are no repairmen working on any planes,

$$1 - 22.69/399.17 = 0.891.$$

3.6 Discussion

A two-stage exact fixed-width confidence interval method was constructed. It was shown that this procedure would have all of the optimal properties asymptotically as the purely sequential asymptotic fixed-width confidence approach. Not only that, but an example was used to show that it does work. The widths of the confidence intervals were just within the bound constraints and both confidence intervals contained the specified parameter. It is well documented that the ratio of two Gamma distributed variables are Beta. This answers the question of operational availability in terms of a 1-a confidence interval; similar research should be performed on a Beta distribution. Mukhopadyay and Zacks (2007), developed a two-stage bounded risk procedure for the Exponential distribution where the parameter of interest was a linear combination of location and scale. Also combining location with scale for a three-parameter Gamma distribution is of interest. This is another area where future research can be performed.

Chapter 4
Conclusion

The goal in this problem was to develop a sampling method that obtained reliable estimators without over-sampling in the Gamma environment. We have found two two-stage sequential methods of finding an appropriate sample size to achieve specified goals: (1) bounding the risk and (2) bounding width of the confidence interval. These methods are both genuine two-stage sampling procedures, meaning it uses information from all observations (initial and additional), and *exact*, meaning only the Gamma distribution was assumed and at no point were there any approximations.

There is mathematical theory supporting the results for when shape is known; the proposed procedures will always yield a reliable estimator. When shape is unknown, it is shown through simulations that inserting an estimator for the shape works nearly as well.

It is also important to realize that the goal is to not simply sample so that the risk and confidence interval widths are within our bounds, but it was desired to sample the fewest number of observations that do so. Result bounded risk results yielded a terminal sample size that was between two and three times the ideal sample size. After investigating the distribution of the risk of the sampling procedure it was found that the bound could be improved. These improved results were giving nearly ideal estimated risks. This gives a more practical usage of the sampling procedure. The interval estimator given yielded nearly ideal results. There was not much room for improvement. The width is always just below d.

Once these methods were constructed they were implemented on two examples: One with real data and the other with simulated data that could be used in a real scenario. The first, observing precipitation intensity of West Point, GA. Forty initial observations were drawn. The assumption was that $\alpha = 2$. We showed how it would affect the sample size and affect the estimate that if $\alpha = 2 \pm 0.25$.

Secondly, we used an aeronautical maintenance example. The operational availability was defined as the ratio of available time over maintenance time plus the available time. The data was simulated according to Morales et al. (2007) and Rodrigues' (1999) paper, which provided a better description of this problem. These procedures are best used when data collection is difficult, expensive, or time-consuming.

Future problems of interest entail: making a more robust estimate with respect to the shape parameter α. As noticed in section 2.4 and section 2.6, if the shape that is assumed known is off by even the smallest margin the result will end in a drastic change in the total number of observations. As mentioned earlier, the Gamma distribution is a flexible right skewed distribution with a positive support. In nature, the support may not necessarily be greater than zero. There exists such a thing as a shift parameter or a truncation parameter that modifies the distribution. So another problem worth looking at is a three-parameter Gamma population. As the example in 3.5 indicated it might be appropriate to extend this research to the Beta population.

Bibliography

[1] Amero, C. and Bayarri, M.J. (1997) *A Bayesian Analysis of a Queueing System with Unlimited Service*, Journal of Statistical Planning and Inference, Vol 58, pp.241-261.

[2] Barber, C. and Jennison, C. (2002),*Optimal Asymmetric One-Sided Group Sequential TestsOptimal Asymmetric One-Sided Group Sequential Tests*. Biometrika, Vol. 89, No. 1, pp. 49-60.

[3] Basawa, I.V. and Prabhu, N. U. (1981), *Estimation in Single Server Queues*, Naval Research Logistics Quarterly, Vol 28, pp. 475-487.

[4] Birnbaum, A. and Healy, W. C., Jr. (1960). *Estimates with Prescribed Variance Based on Two-Stage Sampling*, Annals of Mathematical Statistics, Vol 31, pp. 662-676.

[5] Clarke, A. (1957), *Maximum Likelihood Estimates in a Simple Queue*, Annals of Mathematical Statistics, Vol 28, pp. 1036-1040.

[6] Casella, G and Berger, R. L (2002). Statistical Inference. 2nd Ed. California: Duxbury.

[7] Chapman (1950), *Some Two sample Tests*. Ann Math Statistics, Vol 21, pp 601-606.

[8] Chatelian, F et al. (2007), *Bivariate Gamma Distributions for Image Registration and Change Detection*. IEEE Transactions on Image Processing. Vol 16, No 7, pp.1796-1806.

[9] Chatelian, F et al. (2008), *Change Detection in Multisensor SAR Images Using Bivariate Gamma Distributions*. IEEE Transactions on Image Processing. Vol 17, No 3, pp. 249-258.

[10] Choe, J. and Shroff, N. (1997), *A New Method to Determine Queue Length Distribution at an ATM Multiplexer*, pp. 549-554.

[11] Chow, Y. S. and Robbins, H. (1965). On the asymptotic theory of fixed-width sequential confidence interval for the mean. Ann. Math. statist, 36, 457-462.

[12] Coles, S and Tawn, J. (1996) *A Bayesian Analysis of Extreme Rainfall Data*, Applied Statistics. Vol 45, No 4 , pp. 463-478.

[13] Coit, D. and Jin, T. (1999)*Gamma Distribution Parameter Estimation for Field Reliability Data with Missing Failure Times*, IIE Transactions. Vol 32, pp. 1161-1166.

[14] Dopke, J (1992), *Estimation of Parameters in Gamma Distribution.* IJQRM. pp. 27-43.

[15] Dmitrienko, A. and Govindarajuru, Z. (2000), *Sequential Confidence Regions for Maximum Likelihood Estimates.* The Annals of Statistics , Vol 28, No. 5, pp. 1472-1501.

[16] Finney, D. J. (1984) *Improvement by Planned Multistage Selection* , Journal of the American Statistical Association, Vol 79, No 387, pp. 501- 509.

[17] Gardiner, J and Susarla, V. (1984) *Risk-Efficient Estimation of the Mean Exponential Survival Time under Random Censoring.* Proceedings of the National Academy of Sciences of the United States of America. Vol 81, No 18, Pt 2, pp. 5906-5909.

[18] Ghosh, B.K. and Sen, P.K. (1991) *Handbook on Sequential Analysis.* Marcel Dekker: New York, NY.

[19] Ghosh, B.K. and Muhkopadhyay, N. (1981) *Consistency of Asymptotic Efficiency of Two-Stage and Sequential Procedures.* Sankhya A. Vol 43, pp. 220-227.

[20] Govindarajulu, Z. (1995), *Sequential Point and Interval Estimation of Scale Parameter of an Exponential Distribution* Internat. J. Math. and Sci. Vol 18, No 2, pp. 383-390.

[21] Gutowksi, W.J et al. (2007), *A Possible Constraint on Regional Precipitation Intensity Under Global Warming, Journal of Hydrometerology.* Vol 8, pp.1382-1392.

[22] Hass, Thomas, and Weir (2007) *University Calculus.* New York: Pearson Education.

[23] Hoel, D.B, Sobel, M, and Weiss, G.H., (1972). *A Two-Stage Procedure for Choosing the Better of Two Binomial Populations.* Biometrika. Vol 59, No 2, pp. 317-322.

[24] Kubokawa, T. (1989). *Improving on Two-Stage Estimators for Scale Families*, Metrika, Vol 36, pp. 7-13.

[25] Lorden, G. (1983). *Asymptotic Efficiency of Three-Stage Hypothesis Tests.* The Annals of Statistics. Vol 11, No 1, pp.129-140.

[26] Lui, Q, Proschan, M.A., and Pledger, GW. (2002) *A Unified Theory of Two-Stage Adaptive Designs.* Journal of the American Statistical Association. Vol 97, No 460, pp. 1034-1041.

[27] Maureil, A. et al. (2007) *Impacts of Climate Change on the Frequency and Severity of Floods in the Chateauguay River Basin*, Canada. Can J Civ Eng. Vol 34, pp. 1048-1059.

[28] Morales, J. Castellanosb, M. Mayorala, A. Friedc, R. Armerod, C. (2007)*Bayesian Design in Queues: An Application to Aeronautic Maintenance*, Journal of Statistical Inference, Vol 137, pp. 3058-3067.

[29] Mukhopadyay, N., Silva, B.M., and Waikar, V. (2006), *On a New Two-Stage Confidence Interval Procedure and Comparisons with Its Competitors for Estimating the Difference of Normal Means*

[30] Mukhopadhyay, N. and Pepe, W. (2006) *Exact Bounded Risk Estimation When the Terminal Sample Size and Estimator Are Dependent: The Exponential Case*, Sequential Analysis, Vol 25, No 1, pp. 85 - 101.

[31] Mukhopadhyay, N. (1995). *Two-Stage and Multi-Stage Estimation, in The Exponential Distribution: Theory, Methods and Application*, N. Balakrishnan and A. P. Basu, eds., pp. 429-452, Amsterdam: Gordon and Breach.

[32] Mukhopadyay, N. and Duggan, W. T. (1999). *On a Two-Stage Procedure Having Second-Order Properties with Applications.* Annals Inst. Statistical Mathematics. Vol 51, No 4, pp. 621-636.

[33] Mukhopadhyay, N. and Duggan, W. T. (2000). *New Results on Two-Stage Estimation of the Mean of an Exponential Distribution*, in Perspectives in Statistical Science, A. K. Basu, J. K. Ghosh, P. K. Sen, and B. K. Sinha, eds., pp. 219-231, New Delhi: Oxford University Press.

[34] Phillipi, T. (2005) *Adaptive Cluster Sampling for Estimation of Abundances within Local Populations of Low-Abundance Plants*, Ecology, Vol. 86, No 5, pp. 1091-1100.

[35] Ramsey, F. and Schafer, D. (2002) *The Statistical Sleuth: A Course in Methods of Data Analysis*, 2nd Ed., pp. 102, California: Duxbury

[36] Rasmussen, S. (1980). *A Bayesian Approach to a Problem in Sequential Estimation.* The Annals of Statistics. Vol 8, No 6, pp. 1229-1243.

[37] Rodrigues, M.B., Karpowicz, M., Kang, K., (2000). *A Readiness Analysis for the Argentine Air force and the Brazilian navy A-4 fleet via Consolidated Logistics Support.* In: Joines, J.A., Barton, R.R., Kang, K., Fishwick, P.A. (Eds.), Proceedings of the 2000 Winter Simulation Conference, pp. 1068-1074.

[38] Satagopan, J.M, Venkatramen, E.S, and Begg, C.B. (2002). *Two-Stage Designs for Gene-Disease Association Studies. Biometrics.*Vol 58, No 1,pp. 163-170.

[39] Shapiro, C.P. and Wardrop, R. L., (1980) *Bayesian Sequential Estimation for One-Parameter Exponential Families*, Journal of the American Statistical Association, Vol 372, No 75, pp. 984- 988

[40] Shapiro, C.P. and Wardrop, R. L., (1978) *The Bayes Sequential Procedure for Estimating the Arrival Rate of a Poisson ProcessThe Bayes Sequential Procedure for Estimating the Arrival Rate of a Poisson Process.* Journal of the American Statistical Association, Vol. 73, No. 363, pp. 597-601

[41] Stein, C (1945), *Two-sample Test of a Linear Hypothesis Whose Power is Independent of the Variance.* Ann Math Statistics. Vol 16, pp. 243-258.

[42] Stein, C. (1949), *Some Problems in Sequential Estimation*, Econometrica. Vol 17, pp. 77-78.

[43] Stroud, J.R., Muller, P., and Rosner, G. (2001). *Optimal Sampling Times in Population Pharmacokinetic Studies* .Applied Statistics. Vol 50, No 3, pp. 345-359.

[44] Thall, P.F., Simon, R., and Ellenberg, S. (1988).*Two-Stage Selection and Testing Designs for Comparative Clinical Trials.* Biometrika, Vol 75, No 2, pp. 303-310.

[45] Wald, A (1947) *Sequential Analysis.* Wiley: New York

[46] Whittemore, A. (1997), *Multistage Sampling Designs and Estimating Equations*, Journal of the Royal Statistical Society. Series B, Vol 59, No 3, pp. 589-602.

[47] Yao, T. and Venkatraman, E.S (1998), *Optimal Two-Stage Design for a Series of Pilot Trials of New Agents*, Biometrics, Vol 54, No 3, pp. 1183-1189.

[48] Zacks, S. and Mukhopadhyay, N. (2006a), *Exact Risks of Sequential Point Estimators of the Exponential Parameter*, Sequential Analysis, Vol 25, No 2, pp. 203-226.

[49] Zacks, S. and Mukhopadhyay, N. (2006b) *Bounded Risk Estimation of the Exponential Parameter in a Two-Stage Sampling*, Sequential Analysis. Vol 25, No 4, pp. 437 - 452.

[50] Zacks, S. and Mukhopadhyay, N. (2007), *Bounded Risk Estimation of Linear Combinations of the Location and Scale Parameters in Exponential Distributions under Two-Stage Sampling.* Journal of Statistical Planning and Inference. Vol 137, pp. 3672 - 3686

[51] Zielenzy, M. and Dunn, O.J. (1975), *Cost Evaluation of a Two-Stage Classification Procedure. Biometrics*,Vol 31, No1, pp. 37-47.

Appendix A

Tables

Table A.1: Shape Known, Scale Unknown

n*	w	$\overline{N}$	$\bar{r}$	$\overline{N}$	$\bar{r}$	$\overline{N}$	$\bar{r}$
		m=10		m=20		m=30	
	α=1						
25	2.000	82.966	0.904	64.237	1.011	60.261	0.900
50	1.000	167.069	0.451	129.41	0.565	120.889	0.536
100	0.500	336.415	0.218	255.562	0.248	236.448	0.271
200	0.250	672.591	0.111	509.246	0.128	466.485	0.137
500	0.100	1636.623	0.051	1293.478	0.044	1203.041	0.050
	α=2						
25	1.000	65.287	0.070	57.665	0.614	54.894	0.517
50	0.500	125.489	0.299	114.078	0.252	109.864	0.255
100	0.250	260.783	0.136	224.726	0.126	219.322	0.120
200	0.125	517.502	0.067	430.193	0.062	434.756	0.059
500	0.050	1266.029	0.026	1125.721	0.028	1083.78	0.021
	α=5						
25	0.400	54.445	0.209	53.099	0.193	51.952	0.222
50	0.200	109.525	0.103	105.41	0.103	103.346	0.095
100	0.100	225.068	0.073	209.056	0.052	206.281	0.051
200	0.050	438.813	0.028	423.745	0.027	412.791	0.023
500	0.020	1112.342	0.010	1042.804	0.010	1030.23	0.009
	α=10						
25	0.200	53.273	0.098	51.986	0.106	51.495	0.102
50	0.100	105.208	0.047	103.298	0.046	101.625	0.052
100	0.050	212.180	0.053	206.826	0.024	203.465	0.026
200	0.025	423.050	0.019	408.146	0.012	408.378	0.012
500	0.010	1057.023	0.006	1022.838	0.005	1014.633	0.005

Table A.2: Shape Known, Scale Unknown (Improved)

n*	w	$\overline{N}$	$\overline{N}_{new}$	$\overline{r}$	$\overline{r}_{new}$	$\overline{N}$	$\overline{N}_{new}$	$\overline{r}$	$\overline{r}_{new}$
		m=20				m=30			
	$\alpha = 0.5$								
25	4.000	83.889	52.759	2.161	2.831	71.844	46.502	1.931	2.126
50	2.000	169.721	103.2	1.136	1.772	142.947	88.233	1.145	1.649
100	1.000	336.094	208.133	0.516	0.944	284.756	174.186	0.549	0.932
200	0.500	667.025	415.743	0.261	0.415	561.352	351.080	0.255	0.408
500	0.200	1677.224	1033.184	0.092	0.153	1407.865	862.738	0.095	0.164
	$\alpha = 1$								
25	2.000	64.237	39.387	1.011	1.557	60.261	30.323	0.900	1.638
50	1.000	129.41	77.521	0.565	0.954	120.889	70.861	0.536	0.938
100	0.500	255.562	154.337	0.248	0.443	236.448	140.914	0.271	0.444
200	0.250	509.246	308.944	0.128	0.207	466.485	283.931	0.137	0.221
500	0.100	1293.478	771.913	0.044	0.081	1203.041	704	0.05	0.084
	$\alpha = 2$								
25	1.000	57.665	33.176	0.614	0.858	54.894	34.117	0.517	0.637
50	0.500	109.078	65.551	0.252	0.474	109.864	63.177	0.255	0.455
100	0.250	219.726	130.956	0.126	0.220	219.322	125.347	0.12	0.227
200	0.125	430.193	259.412	0.062	0.110	434.756	250.698	0.059	0.111
500	0.050	1125.721	653.746	0.028	0.042	1083.78	625.733	0.021	0.043
	$\alpha = 5$								
25	0.400	53.099	29.181	0.193	0.359	51.952	31.283	0.222	0.302
50	0.200	105.41	57.886	0.103	0.187	103.346	56.869	0.095	0.189
100	0.100	211.056	115.405	0.052	0.095	206.281	113.396	0.051	0.093
200	0.050	404.745	229.763	0.027	0.046	412.791	226.197	0.023	0.045
500	0.020	1025.804	573.207	0.010	0.018	1030.23	565.884	0.009	0.018
	$\alpha = 10$								
25	0.200	31.986	27.359	0.106	0.191	51.495	30.306	0.102	0.158
50	0.100	82.165	54.219	0.046	0.094	101.625	53.797	0.052	0.099
100	0.050	185.08	107.75	0.024	0.047	203.465	107.384	0.026	0.047
200	0.025	391.815	215.353	0.012	0.024	408.378	213.713	0.012	0.023
500	0.010	997.817	536.765	0.005	0.009	1014.633	533.394	0.005	0.095

Table A.3: Old Bound B as function of Shape and Initial Sample Size

α	m=5	10	15	20	25	30	35	40
1	5.0000	3.0556	2.6374	2.4561	2.3551	2.2906	2.2460	2.2132
2	3.0556	2.4561	2.2906	2.2132	2.1684	2.1391	2.1185	2.1032
3	2.6374	2.2906	2.1882	2.1391	2.1103	2.0914	2.0780	2.0681
4	2.4561	2.2132	2.1391	2.1032	2.0820	2.0681	2.0582	2.0508
5	2.3551	2.1684	2.1103	2.0820	2.0653	2.0542	2.0464	2.0405
6	2.2906	2.1391	2.0914	2.0681	2.0542	2.0451	2.0386	2.0337
7	2.2460	2.1185	2.0780	2.0582	2.0464	2.0386	2.0330	2.0288
8	2.2132	2.1032	2.0681	2.0508	2.0405	2.0337	2.0288	2.0252
9	2.1882	2.0914	2.0604	2.0451	2.0360	2.0299	2.0256	2.0224
10	2.1684	2.0820	2.0542	2.0405	2.0323	2.0269	2.0230	2.0201
20	2.0820	2.0405	2.0269	2.0201	2.0161	2.0134	2.0115	2.0100
30	2.0542	2.0269	2.0179	2.0134	2.0107	2.0089	2.0076	2.0067
40	2.0405	2.0201	2.0134	2.0100	2.0080	2.0067	2.0057	2.0050
50	2.0323	2.0161	2.0107	2.0080	2.0064	2.0053	2.0046	2.0040

Table A.4: New Bound B as function of Shape and Initial Sample Size

α	m=5	10	15	20	25	30	35	40
1	2.8250	1.7264	1.4901	1.3877	1.3306	1.2942	1.2690	1.2505
2	1.7585	1.4233	1.3424	1.3011	1.2768	1.2608	1.2495	1.2411
3	1.4847	1.3394	1.2865	1.2603	1.2447	1.2344	1.2270	1.2215
4	1.3607	1.3001	1.2596	1.2406	1.2290	1.2214	1.2160	1.2119
5	1.2885	1.2775	1.2438	1.2289	1.2198	1.2137	1.2094	1.2062
6	1.2403	1.2628	1.2334	1.2212	1.2136	1.2086	1.2051	1.2024
7	1.2054	1.2525	1.2261	1.2157	1.2093	1.2050	1.2019	1.1996
8	1.1786	1.2449	1.2206	1.2117	1.2060	1.2023	1.1996	1.1976
9	1.1573	1.2391	1.2164	1.2085	1.2035	1.2002	1.1978	1.1961
10	1.1398	1.2345	1.2129	1.2060	1.2015	1.1985	1.1964	1.1948
20	1.0514	1.0305	1.0236	1.0202	1.0181	1.0168	1.0158	1.0151
30	1.0374	1.0236	1.0190	1.0168	1.0154	1.0145	1.0138	1.0134
40	1.0305	1.0202	1.0168	1.0151	1.0140	1.0134	1.0129	1.0125
50	1.0260	1.0181	1.0154	1.0140	1.0132	1.0127	1.0123	1.0120

Table A.5: Shape Unknown, Scale Unknown

n*	w	$\overline{N}$	$\bar{r}$	$\overline{N}$	$\bar{r}$
		m=20		m=30	
	α=1				
25	2.000	103.242	1.342	79.399	0.946
50	1.000	179.782	0.917	174.554	0.764
100	0.500	402.195	0.548	317.901	0.404
200	0.250	783.764	0.295	641.393	0.275
500	0.100	2107.969	0.118	1605.782	0.094
	α=2				
25	1.000	84.273	0.663	72.422	0.495
50	0.500	182.960	0.440	132.700	0.341
100	0.250	300.200	0.251	289.190	0.214
200	0.125	666.948	0.166	605.137	0.110
500	0.050	1656.452	0.077	1367.942	0.041
	α=5				
25	0.400	67.646	0.258	66.675	0.201
50	0.200	133.945	0.175	135.896	0.137
100	0.100	293.482	0.101	265.121	0.079
200	0.050	581.209	0.053	506.883	0.040
500	0.020	1401.071	0.022	1288.534	0.019
	α=10				
25	0.200	67.735	0.141	64.976	0.103
50	0.100	143.471	0.093	140.865	0.072
100	0.050	272.815	0.045	249.548	0.036
200	0.025	528.532	0.023	469.934	0.017
500	0.010	1375.398	0.009	1262.119	0.007

Table A.6: Shape Unknown, Scale Unknown (Improved)

		$\overline{N}$	$\overline{N}_{new}$	$\overline{r}$	$\overline{r}_{new}$	$\overline{N}$	$\overline{N}_{new}$	$\overline{r}$	$\overline{r}_{new}$
		m=20				m=30			
	$\alpha = 1$								
25	2.000	103.242	65.677	1.342	1.684	79.399	54.918	0.946	1.294
50	1.000	179.782	123.159	0.917	1.311	174.554	96.834	0.764	1.003
100	0.500	402.195	224.467	0.548	0.896	317.901	182.988	0.404	0.724
200	0.250	783.764	461.122	0.295	0.541	641.393	367.541	0.275	0.426
500	0.100	2107.969	1198.197	0.118	0.259	1605.782	915.492	0.094	0.192
	$\alpha = 2$								
25	1.000	84.273	50.422	0.663	0.845	72.422	47.812	0.495	0.662
50	0.500	182.96	94.277	0.440	0.624	132.7	81.602	0.341	0.497
100	0.250	300.2	191.886	0.251	0.401	289.19	157.667	0.214	0.336
200	0.125	666.946	369.306	0.166	0.241	605.137	320.843	0.110	0.202
500	0.050	1656.452	920.473	0.077	0.107	1367.942	780.697	0.041	0.080
	$\alpha = 5$								
25	0.400	67.646	44.417	0.258	0.349	66.675	44.293	0.270	0.284
50	0.200	133.945	82.793	0.175	0.242	135.896	76.545	0.204	0.189
100	0.100	293.482	162.109	0.101	0.156	265.121	146.963	0.133	0.093
200	0.050	581.209	325.339	0.055	0.090	596.883	290.237	0.073	0.045
500	0.020	1401.071	826.946	0.022	0.037	1288.534	727.976	0.028	0.018
	$\alpha = 10$								
25	0.200	67.735	43.665	0.141	0.171	64.976	43.471	0.103	0.138
50	0.100	143.471	80.38	0.093	0.125	140.865	74.078	0.072	0.102
100	0.050	272.815	158.329	0.045	0.073	249.548	142.139	0.036	0.062
200	0.025	528.532	316.642	0.023	0.042	469.934	287.158	0.017	0.034
500	0.010	1375.634	792.345	0.009	0.019	1262.119	715.269	0.007	0.013

Table A.7: Initial Sample Size Considerations Simulations

	shape=2 m	scale=5 $\bar{r}$	$\overline{N}$	n^*	Pct of Additional Obs
w=0.500					
	5	0.430	98.170	50	96.3%
	10	0.454	72.565	50	45.1%
	50	0.398	61.815	50	23.6 %
	100	0.248	100.000	50	100.0%
	500	0.050	500.000	50	900.0%
	15	0.453	68.800	50	37.6 %
	40	0.457	61.989	50	24.0 %
w=0.250					
	5	0.218	195.386	100	95.4 %
	10	0.217	149.118	100	49.1 %
	50	0.225	122.467	100	22.5 %
	100	0.211	118.243	100	18.2 %
	500	0.050	500.000	100	400.0%
	29	0.216	127.335	100	27.3 %
	78	0.222	119.003	100	19.0 %
w=0.125					
	5	0.104	383.234	200	91.6 %
	10	0.108	294.230	200	47.1 %
	50	0.107	240.540	200	20.3 %
	100	0.111	237.984	200	19.0 %
	500	0.049	500.000	200	150.0 %
	57	0.112	240.774	200	20.4 %
	157	0.111	233.967	200	17.0 %

Table A.8: Confidence Interval Simulations of Terminal Sample Size

$w = 5$	d	M	$w = 4$	d	M	$w = 3$	d	M
$\alpha = 1$	10.0	21.3	$\alpha = 1$	10.0	25.6	$\alpha = 1$	10.0	34.9
	5.0	45.3		5.0	47.2		5.0	44.4
	2.5	179		2.5	177.6		2.5	178.4
	1.0	1115.5		1.0	1120.1		1.0	1083.3
$\alpha = 2$	10.0	10.7	$\alpha = 2$	10.0	11.6	$\alpha = 2$	10.0	13.7
	5.0	23.4		5.0	20.6		5.0	17.8
	2.5	91.7		2.5	79.8		2.5	72.6
	1.0	578.2		1.0	505.3		1.0	451.2
$\alpha = 5$	10.0	10	$\alpha = 5$	10.0	10	$\alpha = 5$	10.0	10
	5.0	22		5.0	18		5.0	13
	2.5	87		2.5	70		2.5	52
	1.0	542		1.0	433		1.0	325
$\alpha = 10$	10.0	10	$\alpha = 10$	10.0	10	$\alpha = 10$	10.0	10
	5.0	22		5.0	18		5.0	13
	2.5	87		2.5	70		2.5	52
	1.0	542		1.0	433		1.0	325

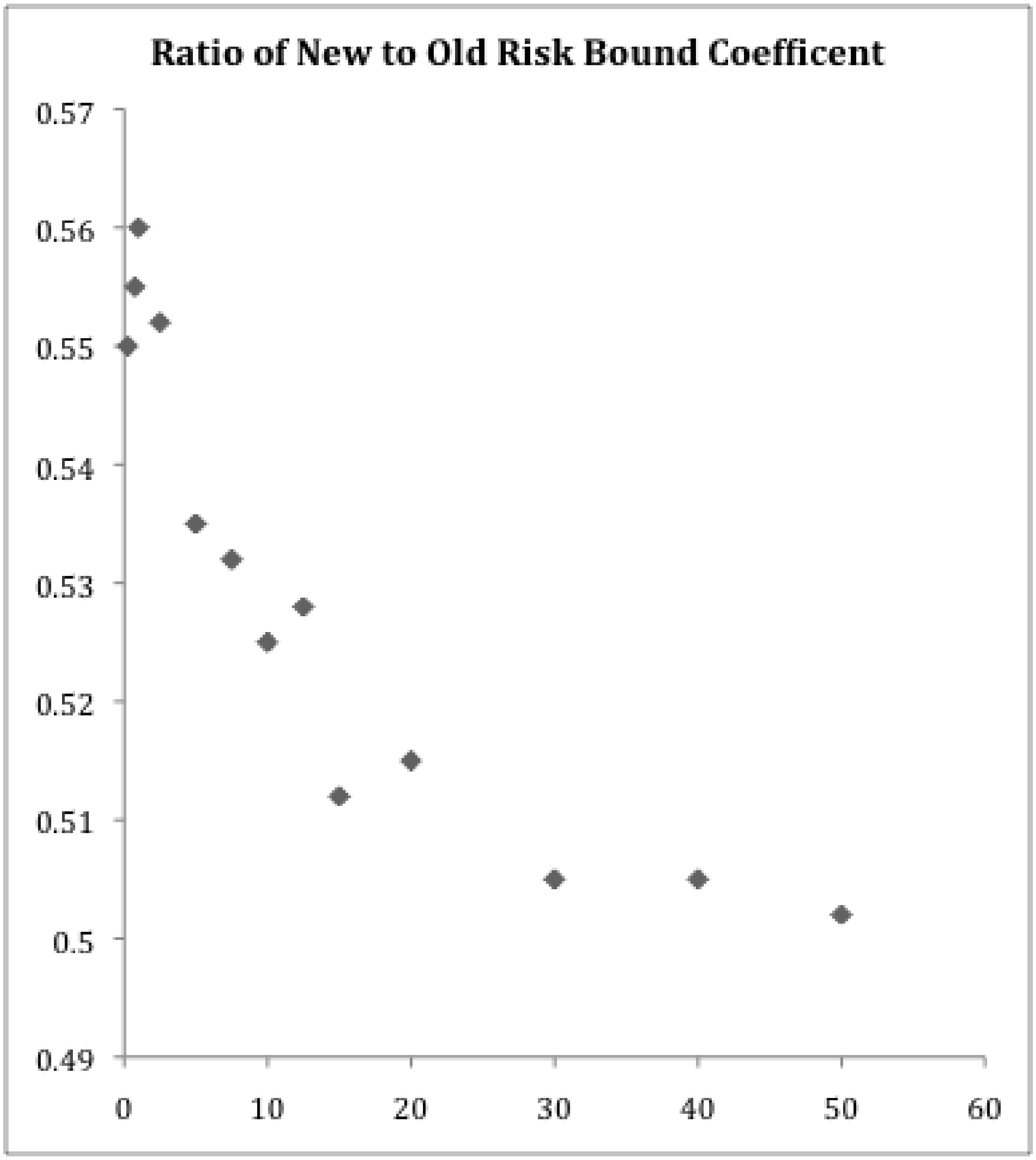
Table A.9: Coverage Percents as Risk Bound Varies

	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$	$\alpha = 5$
	80%			
w=5	0.922	0.956	0.983	1.000
w=4	0.894	0.921	0.977	1.000
w=3	0.920	0.894	0.954	0.999
w=2	0.886	0.882	0.943	0.992
w=1	0.888	0.866	0.850	0.927
	85%			
w=5	0.943	0.968	0.997	1.000
w=4	0.950	0.972	0.991	1.000
w=3	0.940	0.944	0.982	0.999
w=2	0.927	0.906	0.956	0.995
w=1	0.921	0.916	0.875	0.968
	90%			
w=5	0.966	0.987	1.000	1.000
w=4	0.966	0.982	0.998	1.000
w=3	0.958	0.973	0.991	1.000
w=2	0.946	0.955	0.974	0.999
w=1	0.941	0.922	0.921	0.987
	95%			
w=5	0.983	1.000	1.000	1.000
w=4	0.986	0.993	1.000	1.000
w=3	0.970	0.988	0.999	1.000
w=2	0.975	0.970	0.994	0.999
w=1	0.964	0.952	0.962	0.987

Table A.10: Coverage Percents as Initial Sample Size Varies

	m=4	m=6	m=8	m=10	m=12
w=5	0.979	0.994	0.995	0.998	0.999
w=2	0.936	0.954	0.954	0.973	0.984
w=1	0.928	0.934	0.926	0.933	0.938
w=0.5	0.923	0.901	0.901	0.900	0.918
w=0.2	0.921	0.925	0.928	0.924	0.916
w=0.1	0.936	0.914	0.930	0.918	0.927

Appendix B
Figures



As the shape increases the bound coefficient decreases leading to a smaller ratio of old bound coefficient to new.

Figure B.1: Scatterplot of Alpha vs. Empirical Ratios

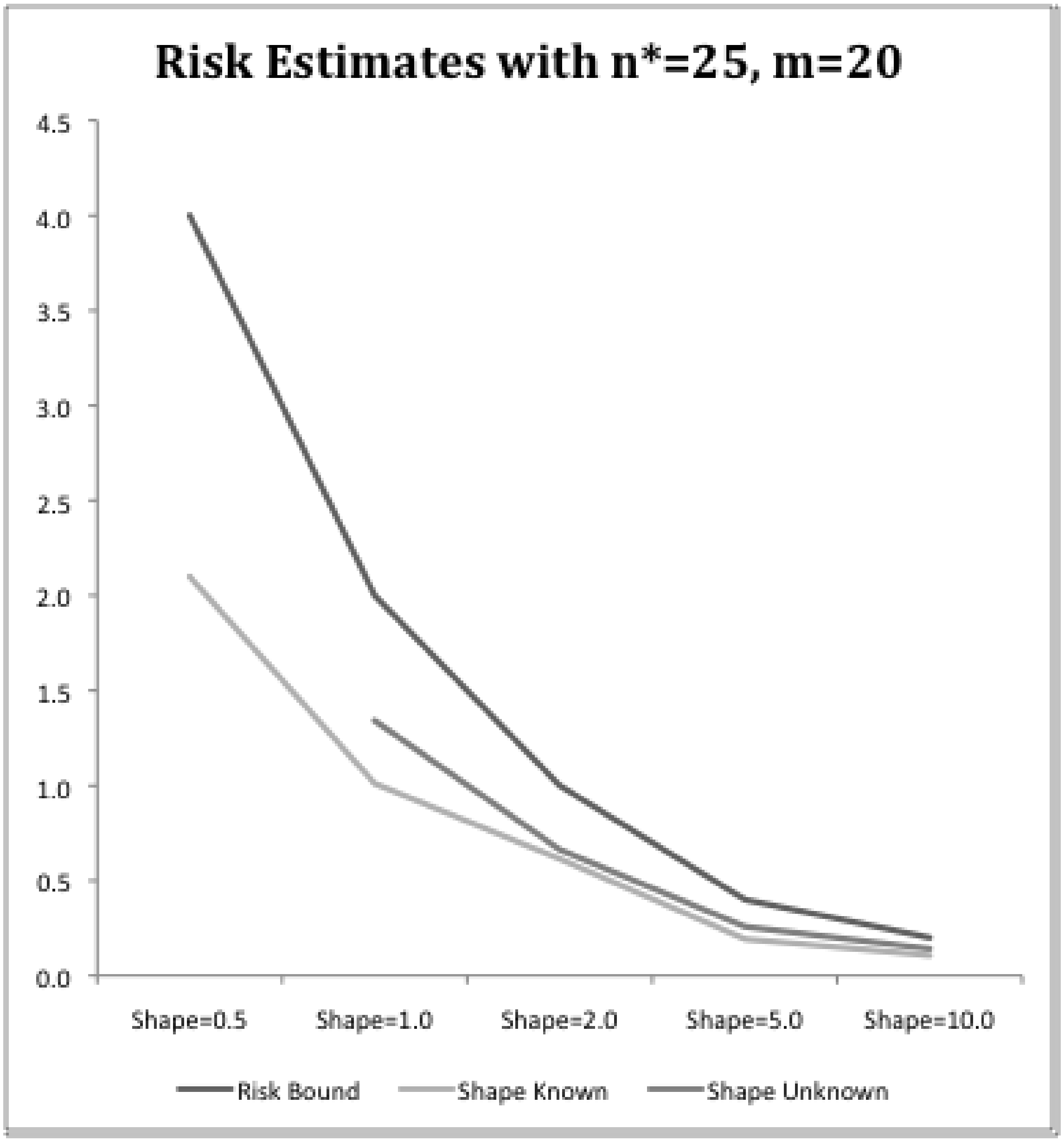


Figure B.2: Average Risk of Two Methods Compared to Risk Bound with Initial Sample of 20

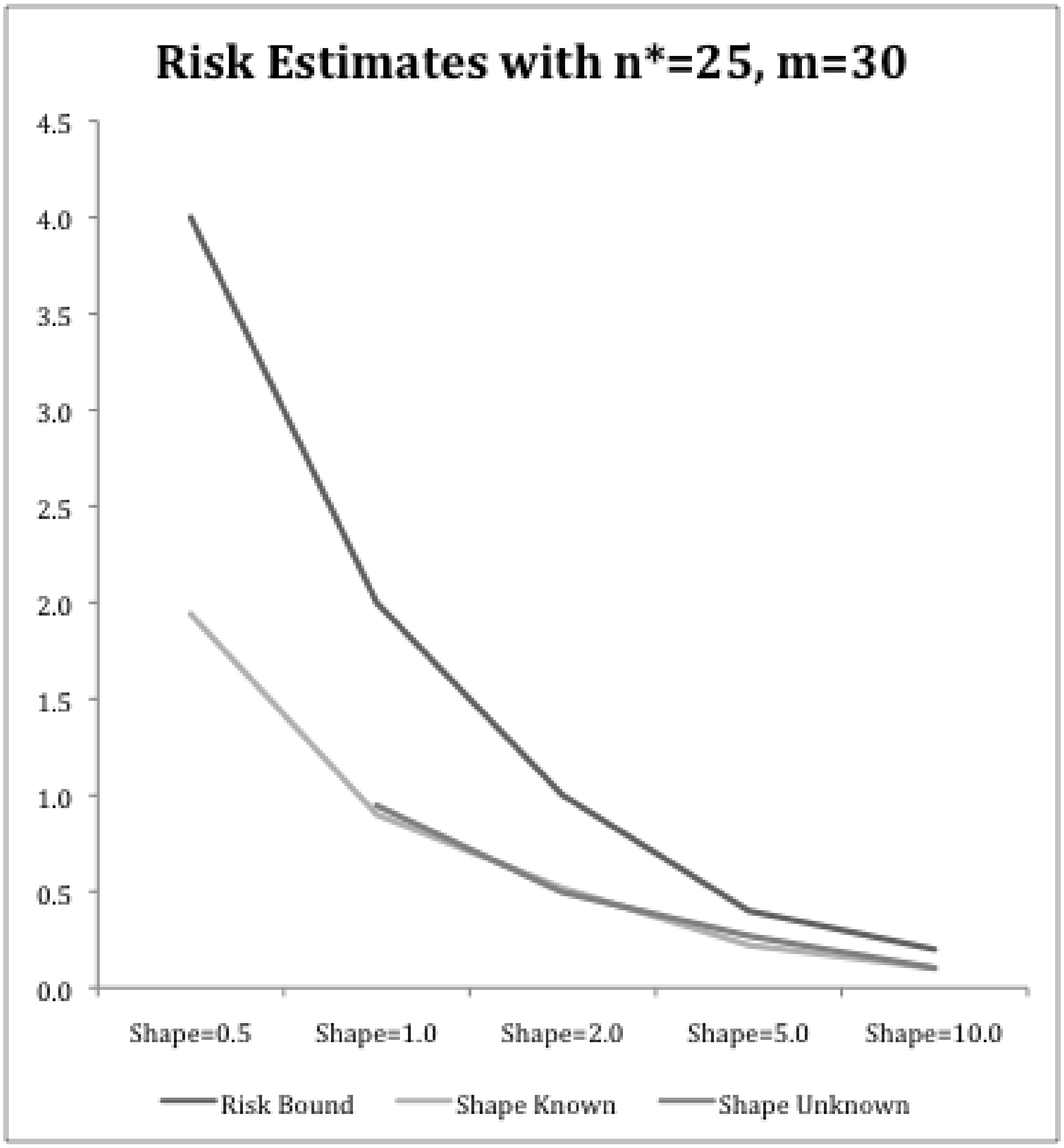


Figure B.3: Average Risk of Two Methods Compared to Risk Bound with Initial Sample of 30

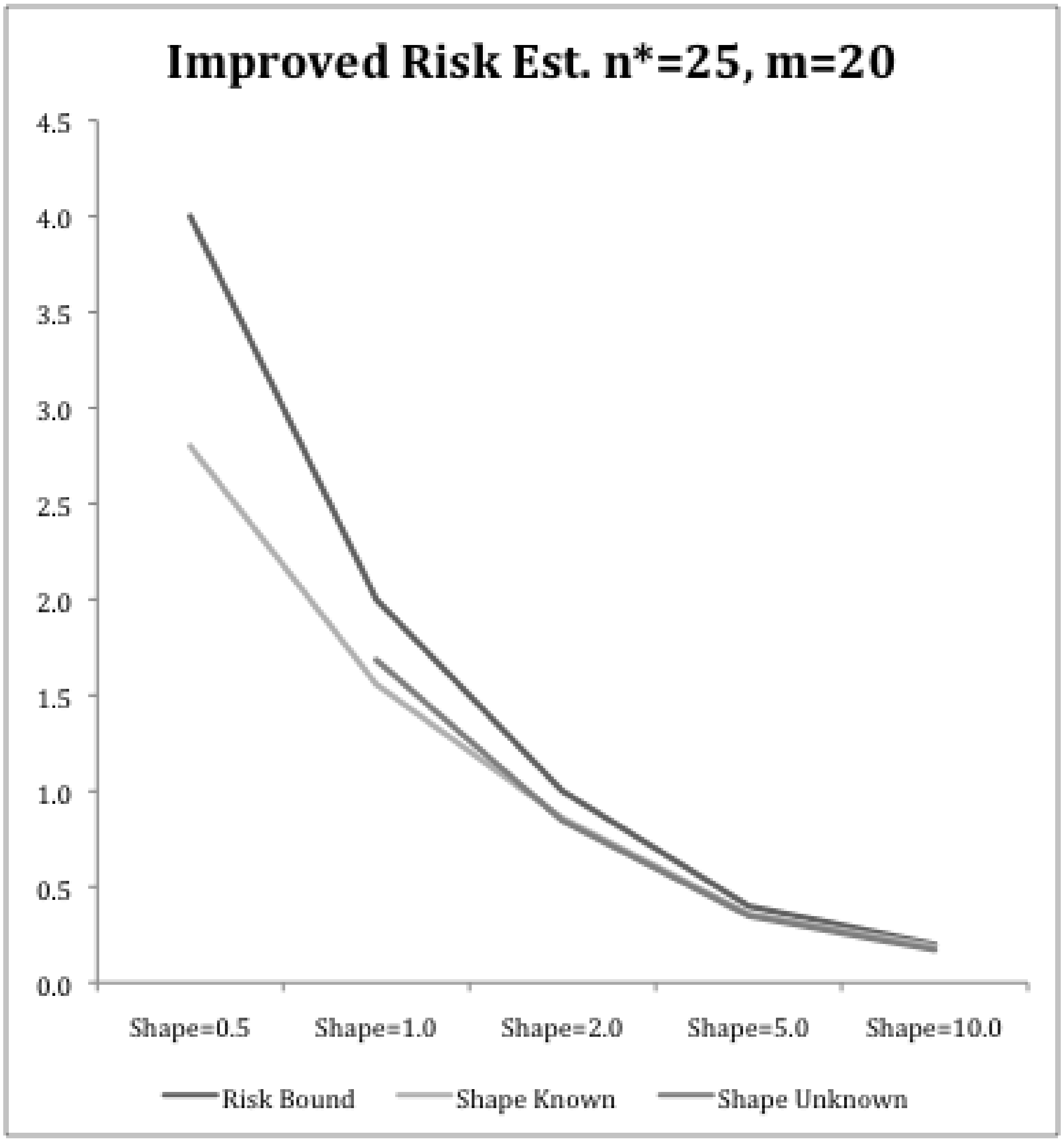


Figure B.4: Average Improved Risk of Two Methods Compared to Risk Bound with Initial Sample of 20

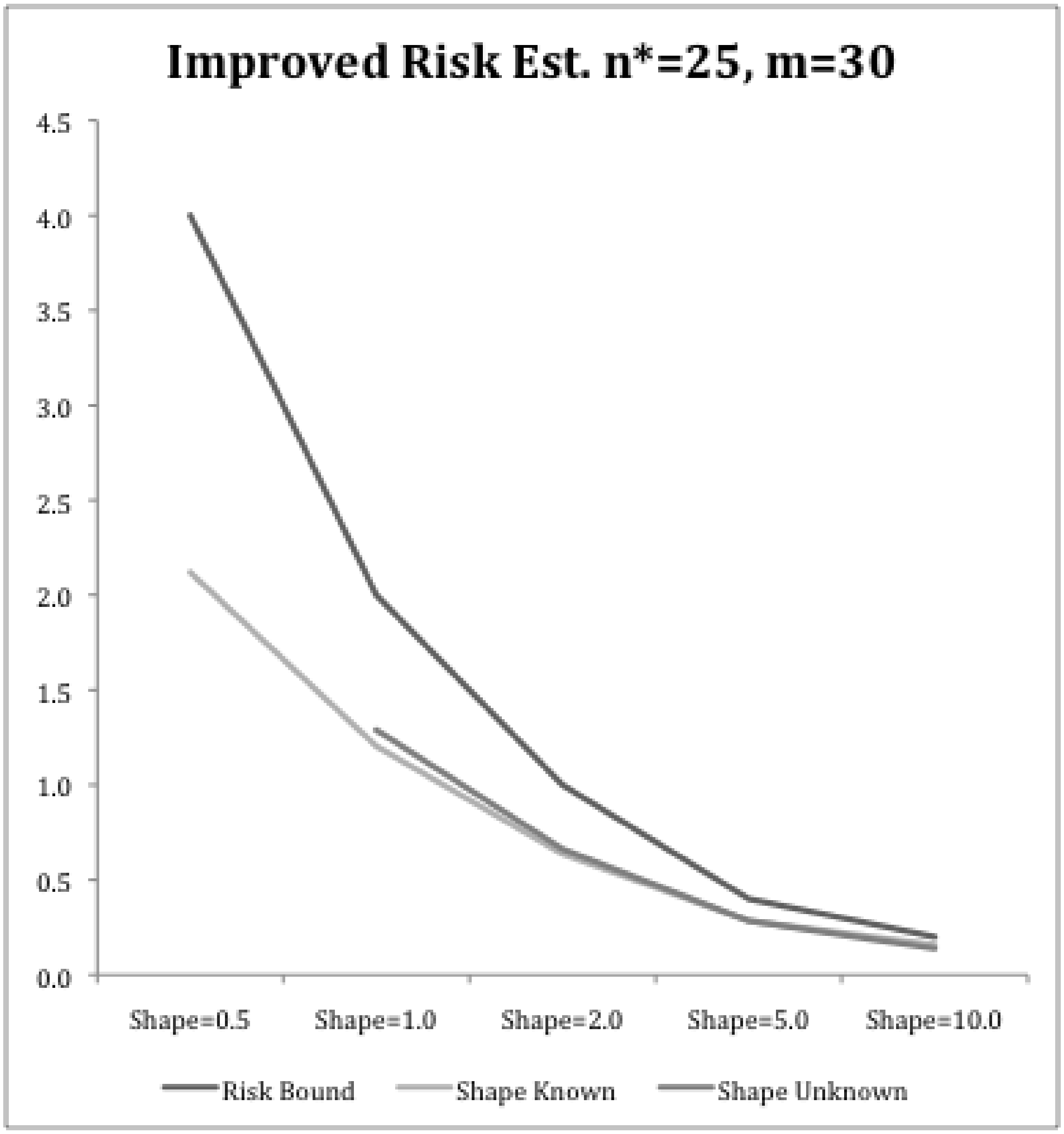


Figure B.5: Average Improved Risk of Two Methods Compared to Risk Bound with Initial Sample of 30